Mechanism Design for LLM Fine-tuning with Multiple Reward Models

Haoran Sun
Peking University
sunhaoran0301@stu.pku.edu.cn

Yurong Chen
INRIA Paris
yurong.chen@inria.fr

Siwei Wang
Microsoft Research Asia
siweiwang@microsoft.com

Wei Chen
Microsoft Research Asia
weic@microsoft.com

Xiaotie Deng
Peking University
xiaotie@pku.edu.cn

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, which could affect the fine-tuned outcomes. 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 a representative class of training rules that maximize social welfare subject to certain regularizations, referred to as SW-Max training rules. First, we show that under most circumstances, truthful reporting is sub-optimal with simply a training rule, thereby highlighting the necessity of payments. Second, we design affine maximizer payment rules that implement SW-Max training rules in dominant-strategy incentive compatibility (DSIC). Further, we characterize sufficient conditions for payment equivalence properties. For a training rule that satisfies these conditions, we have found all the payment rules that implement it in DSIC, as they only differ by a constant term irrelevant to agents' reports from each other.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF, Ouyang et al. [2022], Christiano et al. [2017]) has emerged as a mainstream approach to align Large Language Models (LLMs) with human values. However, the implementation of standard RLHF is often resource-intensive. Constraints such as budget limitations and privacy concerns prevent individuals from obtaining fine-tuned models aligned with their preferences. Consequently, integrating multiple preferences within a single RLHF process becomes valuable but presents practical challenges. From an algorithmic perspective, Multiple-Objective RLHF (MORLHF, Bai et al. [2022], Wu et al. [2024]) offers a promising solution. Following MORLHF, there is further research focusing on improving the efficiency (Rame et al. [2024], Shi et al. [2024], Jang et al. [2023]), accuracy (Eisenstein et al. [2023], Coste et al. [2023], Zhang et al. [2024a], Ramé et al. [2024]), and fairness (Chakraborty et al. [2024]) of algorithms that integrate multiple preferences.

As these techniques advance, it is natural to consider such a potential economic scenario: a platform provides a fine-tuning service to aggregate preferences, and different groups report preferences to the platform on behalf of their agents who share the same preference. We illustrate this by a simplified RLHF scenario presented in Figure 1. As is shown in the figure, there are two groups reporting their preferences, and the model is fine-tuned according to the training rule. Notably, although group 1's

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

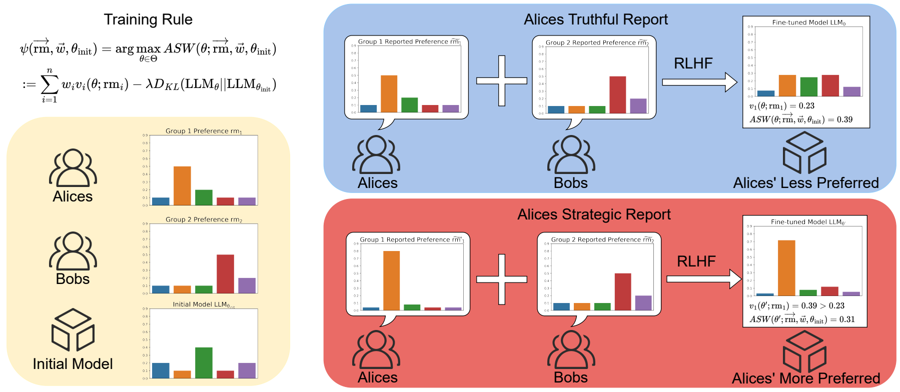

Figure 1: Motivating example of the RLHF Game: Consider a basic training objective ψ in RLHF for two groups, setting group sizes $w_1 = w_2 = 1$. When there is no payment rule and group 2's report $\widetilde{\text{rm}}_2$ is fixed, group 1 can gain a higher utility by strategically reporting $\widetilde{\text{rm}}'_1 \neq \text{rm}_1$ than truthfully reporting.

true preference is rm_1, adopting a reporting strategy $\widetilde{\text{rm}}'_1 \neq \text{rm}_1$ can get a more preferred model than truthfully reporting. In practice, we have to account for the possibility that rational groups will strategically report to maximize their utilities like that. If the mechanism is vulnerable to such manipulation, the final model can be unpredictable, affecting the training outcome. Hence, from the platform's view, it is essential to consider not only the fine-tuning algorithm's performance in achieving specific objectives but also the incentive compatibility of preference reporting.

In this paper, we address the incentive issues for preference reporting within this scenario by framing it scenario as a multi-parameter mechanism design problem and term it the *RLHF Game*. Based on the RLHF Game, *our findings show that many commonly used training objectives lead to profitable misreporting strategies. However, we demonstrate that a simple charging scheme can incentivize truthful reporting, and under certain conditions, this scheme is uniquely determined.*

The rest of the paper is organized as follows: In Section 2, we introduce the formal RLHF Game model. We analyze the incentives within the RLHF Game in Section 3. Further empirical study and related work are provided in Appendix C and Appendix D, respectively.

2 Formulation of the RLHF Game

In this section, we present the formal description of the RLHF Game. In the RLHF Game, there is one LLM *provider* and n *groups of agents*, denoted by $[n] = \{1, 2, \cdots, n\}$. Let $T^* := \emptyset \cup T \cup T^2 \cup \cdots \cup T^K$ represent the set of all possible input sequences with lengths up to K. The provider has an initial model $\text{LLM}_{\theta_{\text{init}}}$ with non-zero 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 and a joint preference represented by a reward model $\text{rm}_i : T^* \to \mathbb{R}_{\geq 0}$. We mainly consider two types of reward models: normalized by summation ($\sum_{x\in T^*} \text{rm}(x) = 1$) and normalized by maximum ($\max_{x\in T^*} \text{rm}(x) = 1$). Let $\mathcal{R}$ and $\mathcal{W} \subseteq \mathbb{N}_+$ denote the domains for each group's reward model and group size, respectively. We assume an upper bound $\bar{w}$ for $\mathcal{W}$. The exact reward model and the size are group i's private information. For an agent in group i, the valuation it receives from a model LLM_θ is the expected reward on the sequences generated by LLM_θ. Formally, for all $i \in [n]$, $v_i(\theta; \text{rm}) = \mathbb{E}_{\boldsymbol{x}\sim \text{LLM}_\theta} \text{rm}(\boldsymbol{x}) = \sum_{\boldsymbol{x}\in T^*} \text{LLM}_\theta(\boldsymbol{x})\text{rm}(\boldsymbol{x})$.

The provider first announces the mechanism, including a training rule $\psi : \mathcal{R}^n \times \mathcal{W}^n \times \Theta \to \Theta$ and a payment rule $p : \mathcal{R}^n \times \mathcal{W}^n \times \Theta \to \mathbb{R}^n$. 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. Specifically, the training rule seeks the model that maximizes a specific objective function OBJ. That is, $\psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) \in \arg\max_{\theta\in\Theta} \text{OBJ}(\theta; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$ (We break the tie based on the further ordering on the $v_i(\theta; \text{rm}_i)$s).

After observing the announced mechanism (ψ, p), each group i reports a reward model, $\widetilde{\text{rm}}_i$, and its group size $\tilde{w}_i$. We assume all reported sizes are in $\mathcal{W}$ and therefore bounded by $\bar{w}$. Based on the reported information, the provider fine-tunes the model and gets the final model with parameter $\theta_{\text{final}} = \psi(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$. The provider then charges group i according to the payment rule, $p_i(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$. All the members in the group have access to the fine-tuned model θ_{final}, so the valuation for group i is $w_i v_i(\theta_{\text{final}}; \text{rm}_i)$. We assume all groups have quasi-linear utilities. Therefore, group i's utility is $u_i(\overrightarrow{\text{rm}}, \vec{w}; \psi, p, \text{rm}_i, w_i) = w_i v_i(\theta_{\text{final}}; \text{rm}_i) - p_i(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$. The groups may strategically report, thus $\overrightarrow{\text{rm}}$ and $\vec{w}$ 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 is accessible to the groups' private information. We formally define these desiderata of a mechanism as follows.

- A mechanism (ψ, p) satisfies *ϵ-dominant-strategy incentive compatibility* (ϵ-DSIC) if $\forall i$, rm_i, w_i, rm'_i, w'_i, $\overrightarrow{\text{rm}}_{-i}$, $\vec{w}_{-i}$, θ_{init}, we have
$$u_i((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i}); \text{rm}_i, w_i) + \epsilon \geq u_i((\text{rm}'_i, \overrightarrow{\text{rm}}_{-i}), (w'_i, \vec{w}_{-i}); \text{rm}_i, w_i). \quad (1)$$
- A mechanism (ψ, p) satisfies *ϵ-individually rationality* (ϵ-IR) if $\forall i$, rm_i, w_i, $\overrightarrow{\text{rm}}_{-i}$, $\vec{w}_{-i}$, θ_{init}, we have
$$u_i((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), (w_i, \vec{w}_{-i}); \text{rm}_i, w_i) + \epsilon \geq 0. \quad (2)$$

In particular, DSIC and IR refer to 0-DSIC and 0-IR, respectively. 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.

3 Incentives in the RLHF Game

This section explores incentive design within the RLHF Game framework. We primarily focus on a representative set of training rules that maximize social welfare with regularization. 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 3.1 (SW-Max Training Rules)**.** A **S**ocial **W**elfare-**Max**imizing training rule (SW-Max training rule) fine-tunes the model to maximize the summation of the groups' valuations subject to a regularization measured by f-divergence (Ali and Silvey [1966], Csiszár [1967], Shi et al. [2024]). Formally, the training objective is $\text{OBJ}(\theta; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) = \sum_{i=1}^n w_i v_i(\theta; \text{rm}_i) - D_f(\text{LLM}_\theta || \text{LLM}_{\theta_{\text{init}}})$, where f is a convex function on $\mathbb{R}_+$ and $f(1) = 0$.

Here, we use $ASW(\theta; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) := \sum_{i=1}^n w_i v_i(\theta; \text{rm}_i) - D_f(\text{LLM}_\theta || \text{LLM}_{\theta_{\text{init}}})$ to denote the affine social welfare and denote $\psi \in \Psi^{SW}$ that training rule ψ belongs to this set.

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 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 3.2. *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. On the other hand, 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. Further, we consider the case that $s_i \geq 2$ and $\underline{\text{rm}_i} = 0$. The analysis in this case is much more complex, and we refer to Theorem A.1.

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 SW-Max training rules, we first define $ASW_{-i}(\theta; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$ the affine social welfare except for group i. That is, $ASW_{-i}(\theta; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) := ASW(\theta; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) - w_i v_i(\theta; \text{rm}_i)$. Then, we derive the affine maximizer payment rule ([Roberts [1979]]) p^{AFF}:

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

We show that p^{AFF} implements SW-Max training rules in both DSIC and IR:

Theorem 3.3. *For any $\psi \in \Psi^{SW}$, mechanism (ψ, p^{AFF}) satisfies DSIC and IR.*

The second question is more general, so we primarily consider the concept of *payment equivalence* ([Ashlagi et al., 2010]) defined as:

Definition 3.4 (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 $\forall \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}$, we have $p'_i(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) = p_i(\overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}}) + g_i\left(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}\right)$.

Payment equivalence indicates that the only way to modify a DSIC mechanism (ψ, p) to (ψ, p') while maintaining incentive compatibility 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.

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. [2007] to show that *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 address this problem, we introduce the following condition that illustrates the continuity of the training rule.

Condition 3.5. 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 $\theta' := \psi(\overrightarrow{\text{rm}}', \vec{w}', \theta_{\text{init}})$.

We also validate this property for some widely used divergence function f in Proposition E.3. Based on this property, we show sufficient conditions of payment equivalence for general training rules.

Theorem 3.6. *An implementable training rule ψ satisfies payment equivalence if Condition 3.5 holds and $\forall i$, $\overrightarrow{rm}_{-i}$, $\vec{w}_{-i}$, 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 $w_i \in \mathcal{W}$. For maximum normalization, that rm_i^* must be $\mathbb{1}$.*

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 $w_i \in \mathcal{W}$. This makes the domain of $\mathcal{R} \times \mathcal{W}$ be connected in another sense that can also induce payment equivalence. 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, meets that criterion. Based on this, we obtain the following result:

Corollary 3.7. *When Condition 3.5 holds, each training rule $\psi \in \Psi^{SW}$ satisfies payment equivalence.*

With Theorem 3.3 and Corollary 3.7, we can conclude that all the payment rules that implement a SW-Max training rule satisfies Condition 3.5 is essential an affine maximize payment rule.

References

Saaket Agashe, Yue Fan, and Xin Eric Wang. Evaluating multi-agent coordination abilities in large language models, 2023.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models, 2023.

Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another, 1966.

Itai Ashlagi, Mark Braverman, Avinatan Hassidim, and Dov Monderer. Monotonicity and implementability, 2010.

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, 2022.

Dirk Bergemann and Juuso Välimäki. The dynamic pivot mechanism, 2010.

Sushil Bikhchandani, Shurojit Chatterji, Ron Lavi, Ahuva Mu'alem, Noam Nisan, and Arunava Sen. Weak monotonicity characterizes deterministic dominant-strategy implementation, 2006.

Patrick Briest, Shuchi Chawla, Robert Kleinberg, and S Matthew Weinberg. Pricing randomized allocations, 2010.

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, 2024.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of gpt, 2023.

Alessandro Chiesa, Silvio Micali, and Zeyuan Allen Zhu. Mechanism design with approximate valuations, 2012.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2017.

Vincent Conitzer and Tuomas Sandholm. Self-interested automated mechanism design and implications for optimal combinatorial auctions, 2004.

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, 2024.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization, 2023.

Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations, 1967.

Michael Curry, Tuomas Sandholm, and John Dickerson. Differentiable economics for randomized affine maximizer auctions, 2022.

Zvi Drezner and Horst W Hamacher. Facility location: applications and theory, 2004.

Zhijian Duan, Haoran Sun, Yurong Chen, and Xiaotie Deng. A scalable neural network for dsic affine maximizer auction design, 2024a.

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, 2024b.

Kumar Avinava Dubey, Zhe Feng, Rahul Kidambi, Aranyak Mehta, and Di Wang. Auctions with llm summaries, 2024.

Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models, 2023.

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, 2023.

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, 2022.

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. Can large language models serve as rational players in game theory? a systematic analysis, 2023.

Soheil Feizi, MohammadTaghi Hajiaghayi, Keivan Rezaei, and Suho Shin. Online advertisements with llms: Opportunities and challenges, 2023.

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, 2024.

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, 2024.

Kanishk Gandhi, Dorsa Sadigh, and Noah D Goodman. Strategic reasoning with language models, 2023.

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, 2024.

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, 2023.

Shangmin Guo, Haoran Bu, Haochuan Wang, Yi Ren, Dianbo Sui, Yuming Shang, and Siting Lu. Economics arena for large language models, 2024a.

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, 2024b.

Birgit Heydenreich, Rudolf Müller, Marc Uetz, and Rakesh V Vohra. Characterization of revenue equivalence, 2009.

Radosveta Ivanova-Stenzel and Timothy C Salmon. Revenue equivalence revisited, 2008.

Athul Paul Jacob, Yikang Shen, Gabriele Farina, and Jacob Andreas. The consensus game: Language model generation via equilibrium search, 2023.

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, 2023.

Philippe Jehiel, Moritz Meyer-Ter-Vehn, and Benny Moldovanu. Mixed bundling auctions, 2007.

Benjamin Laufer, Jon Kleinberg, and Hoda Heidari. Fine-tuning games: Bargaining and adaptation for general-purpose models, 2023.

Anton Likhodedov and Tuomas Sandholm. Methods for boosting revenue in combinatorial auctions, 2004.

Nunzio Lorè and Babak Heydari. Strategic behavior of large language models: Game structure vs. contextual framing, 2023.

David G Luenberger, Yinyu Ye, et al. Linear and nonlinear programming, 1984.

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, 2023.

Mitsunobu Miyake. On the incentive properties of multi-item auctions, 1998.

Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. Welfare diplomacy: Benchmarking language model cooperation, 2023.

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, 2023.

Roger B Myerson. Incentive compatibility and the bargaining problem, 1979.

Roger B Myerson. Optimal auction design, 1981.

Noam Nisan and Amir Ronen. Algorithmic mechanism design, 1999.

Noam Nisan et al. Introduction to mechanism design (for computer scientists), 2007.

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, 2022.

Susan Hesse Owen and Mark S Daskin. Strategic facility location: A review, 1998.

Chanwoo Park, Mingyang Liu, Kaiqing Zhang, and Asuman Ozdaglar. Principled rlhf from heterogeneous feedback via personalization and preference aggregation, 2024.

Alessandro Pavan, Ilya Segal, and Juuso Toikka. Dynamic mechanism design: A myersonian approach, 2014.

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, 2024.

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

Kevin Roberts. The characterization of implementable choice rules, 1979.

Jean-Charles Rochet. A necessary and sufficient condition for rationalizability in a quasi-linear context, 1987.

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, 2024.

Michael Saks and Lan Yu. Weak monotonicity suffices for truthfulness on convex domains, 2005.

Tuomas Sandholm and Anton Likhodedov. Automated design of revenue-maximizing combinatorial auctions, 2015.

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

Ruizhe Shi, Yifang Chen, Yushi Hu, ALisa Liu, Noah Smith, Hannaneh Hajishirzi, and Simon Du. Decoding-time language model alignment with multiple objectives, 2024.

Ermis Soumalias, Michael J Curry, and Sven Seuken. Truthful aggregation of llms with an application to online advertising, 2024.

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, 2020.

Pingzhong Tang and Tuomas Sandholm. Mixed-bundling auctions with reserve prices., 2012.

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, 2023.

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, 2024.

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, 2023.

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, 2024.

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, 2023a.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game, 2023b.

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, 2024.

Shun Zhang, Zhenfang Chen, Sunli Chen, Yikang Shen, Zhiqing Sun, and Chuang Gan. Improving reinforcement learning from human feedback with efficient reward model ensemble, 2024a.

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, 2024b.

A Necessity of Payment Rule

We consider the case that $s_i \geq 2$ and $\underline{\text{rm}_i} = 0$. Since the minimum value is already 0, the strategy demonstrated in Theorem 3.2 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 change in the valuation for group i when the reported value $\text{rm}_i(\boldsymbol{x})$ is slightly adjusted from its truthful value. 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. We summarize this in the following theorem, but due to the complicated form of the function t, we provide the detailed explanation in the Theorem E.2.

Theorem A.1. *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.*

Combining Theorem 3.2 and Theorem A.1, 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 failure of a mechanism without payments to incentivize truthfulness.

B Approximate Valuation Model

In this part, we discuss the influence of error generated in practice on the incentive property in the RLHF Game. We abstract it as an approximate valuation problem (Chiesa et al. [2012]). Formally, when group i reports its reward model rm_i, the mechanism may not be accessible to rm_i but rather a noisy reward model $\widehat{\text{rm}}_i$ with a conditional distribution $F_i(\cdot|\text{rm}_i)$, and then use it as the input. We use

$F(\cdot|\overrightarrow{\text{rm}})$ to denote the joint distribution of these independent distributions. This abstraction can model various errors that may occur in practical training. One example is that the calculation of valuation defined in Section 2 requires sampling sequences from LLM, which may result in a deviation from the true valuation.

In this model, we assume that groups are aware of the noise when feeding preferences into the mechanism. Therefore, their utilities will take it into account and have a different form. We use the capital letter U_i to represent agent i's revised utility. Formally, for group i with reward model rm_i and group size w_i, its utility for reporting (rm'_i, w'_i) is 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 when the noised input reward models to the mechanism and the reported reward models are close:

Condition B.1. For any profile of reported reward models $\overrightarrow{\text{rm}}$, the reward models $\overrightarrow{\widehat{\text{rm}}}$ that can be generated from $F(\cdot|\overrightarrow{\text{rm}})$ with non-zero probability satisfies

$$\max_{\boldsymbol{x} \in T^*} |\widehat{\text{rm}}_i(\boldsymbol{x}) - \text{rm}_i(\boldsymbol{x})| \leq \epsilon \quad \forall i \in [n].$$

We explore the influence of such errors on both the training objective and the incentive compatibility. Firstly, we show that by directly applying mechanism in Section 3 to the noised input, the loss in the social welfare is upper-bounded by $2\epsilon \sum_{i=1}^{n} w_i$.

Lemma B.2. *When Condition B.1 holds and the training rule $\psi \in \Psi^{SW}$, if all groups truthfully report, the loss in social welfare is bounded by $2\epsilon \sum_{i=1}^{n} w_i$.*

For training rule $\psi \in \Psi^{SW}$, a group's utility in the mechanism (ψ, p^{AFF}) consists of an affine social welfare term ASW. Therefore, we can derive the following theorem based on Lemma B.2.

Theorem B.3. *When Condition B.1 holds and the training rule $\psi \in \Psi^{SW}$, for group i and any rm_i, rm'_i, $\overrightarrow{rm}_{-i}$, w_i and $\vec{w}_i$, we have*

$$U_i((rm_i, \overrightarrow{rm}_{-i}), (w_i, \vec{w}_{-i}); \psi, p^{AFF}, rm_i, w_i) \geq \\ U_i((rm'_i, \overrightarrow{rm}_{-i}), (w_i, \vec{w}_{-i}); \psi, p^{AFF}, rm_i, w_i) - 2w_i\epsilon.$$

This theorem implies that when we do not consider the strategic report for $\vec{w}$, the mechanism (ψ, p^{AFF}) satisfies $\max_{i \in [n]} 2w_i\epsilon$-DSIC. 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 is costlier than $2w_i\epsilon$.

C Empirical Study

In this section, we present an empirical evaluation of the mechanism, focusing on the DSIC property and illustrating how payment rules incentivize truthful reporting in practical applications. The experimental results demonstrate the effectiveness of our proposed mechanism in real-world training scenarios.

C.1 Setup

Our experimental setup mainly follows the literature that studies MORLHF (Wu et al. [2024]) and the improved method for multiple objectives training for LLMs, like Rewarded Soups (Rame et al. [2024]), Rewards-in-Context (Yang et al. [2024]), and Multi-Objective Decoding (Shi et al. [2024]). We consider two tasks: the Helpful Assistants task (Bai et al. [2022]) and the Reddit Summary task (Stiennon et al. [2020]). And we use LLAMA2-7B (Touvron et al. [2023]) as the base model for both tasks.

We get the initial model $\text{LLM}_{\theta_{\text{init}}}$ for the Helpful Assistants task by first supervised fine-tuning an LLAMA2-7B model on the Anthropic-HH dataset (Bai et al. [2022]). Then, we use two reward models that measure harmlessness and humor for the RLHF process, respectively. For the Reddit Summary task, the supervised fine-tuning is on the Summarize-from-Feedback dataset (Stiennon et al. [2020]). We use two reward models for this task, measuring the summary's quality and faithfulness.

We formulate these tasks as two RLHF games: a "Harmless vs. Humor" game for the Helpful Assistants task and a "Faithful vs. Summary" game for the Reddit Summary task. In each game, the reward models represent the true preferences of two distinct groups: for instance, in "Harmless v.s. Humor," group 1 prioritizes harmlessness while group 2 values humor. We denote the reward models for these preferences as rm_1 (harmlessness) and rm_2 (humor), with group size vectors (w_1, w_2) selected from $\{(3,7),(5,5),(7,3)\}$, varying across different settings.

C.2 Implementation Details

We implement the basic training rule described in Definition 3.1 and use KL-divergence as the distance measure f. We first train models using single reward models and then combine them using the technique of Rewarded Soups (Rame et al. [2024]) and Multi-Objective Decoding (Shi et al. [2024]) to produce a set of hybrid models $\{\theta_1, \theta_2, \cdots, \theta_K\}$. These hybrid models form the set Θ in Definition 3.1. As demonstrated in Rame et al. [2024], Shi et al. [2024], this method reduces training costs while maintaining results comparable to full multi-objective fine-tuning.

Given the large space for potential misreporting, we consider two simple misreporting strategies, $(\widetilde{\text{rm}}_i, \tilde{w}_i)$ for group i while keeping the other group's report fixed:

1. $\widetilde{\text{rm}}_i = \text{rm}_i$ and $\tilde{w}_i = \alpha w_i$,
2. $\widetilde{\text{rm}}_i = \beta \text{rm}_i + (1-\beta)\text{rm}_{-i}$ and $\tilde{w}_i = w_i$.

Strategy (1) involves misreporting the group size, and strategy (2) leverages the other group's information to misreport preferences. Intuitively, by adopting a larger α or β, group i can gain more influence in the training process and ultimately obtain a fine-tuned model it prefers more. Among these, $\alpha = 1$ and $\beta = 1$ represent truthfully reporting.

C.3 Result Analysis

Since the values output by different reward models have varying scales, we normalize all reward values to $[0, 1]$, ensuring that the maximum and minimum values are 1 and 0, respectively. We then report the valuations, payments, and utilities of group i for different reporting strategies under the mechanism, based on the normalized values, in Figure 2. Each column represents a specific group size (w_1, w_2), with the first three columns corresponding to the "Harmless vs. Humor" task and the last column to the "Faithful vs. Summary" task.

As illustrated in the figure, increasing the parameter α (or β) leads to a higher valuation for the group, revealing the failure of non-payment mechanisms to incentivize truthful reporting. However, when payments are computed according to p^{AFF}, the payment increases alongside α or β, balancing the group's valuation and ensuring that truthful reporting ($\alpha = 1$, $\beta = 1$) maximizes utility in all cases.

D Related Work

D.1 Primary Related Work

Several studies have investigated similar scenarios. Among them, Duetting et al. [2023], Soumalias et al. [2024] and Park et al. [2024] are most related to ours. Duetting et al. [2023] examines the problem of designing a mechanism to aggregate multiple agents' preferences based on each agent's bids and determine their payments. However, they exclude the case where preferences can be misreported, which is the primary concern in our study. The concurrent work of Soumalias et al. [2024] and Park et al. [2024] also considers the mechanism design for strategic preference reporting behavior. However, Soumalias et al. [2024] mainly focuses on the practical implementation of SW-Max training rule with KL-divergence and the payment scheme. And Park et al. [2024] primarily discusses the implementability of a training rule. In this work, we provide a more comprehensive analysis of strategic reporting and are concerned with the theoretical properties of more general mechanisms, including implementability and payment equivalence.

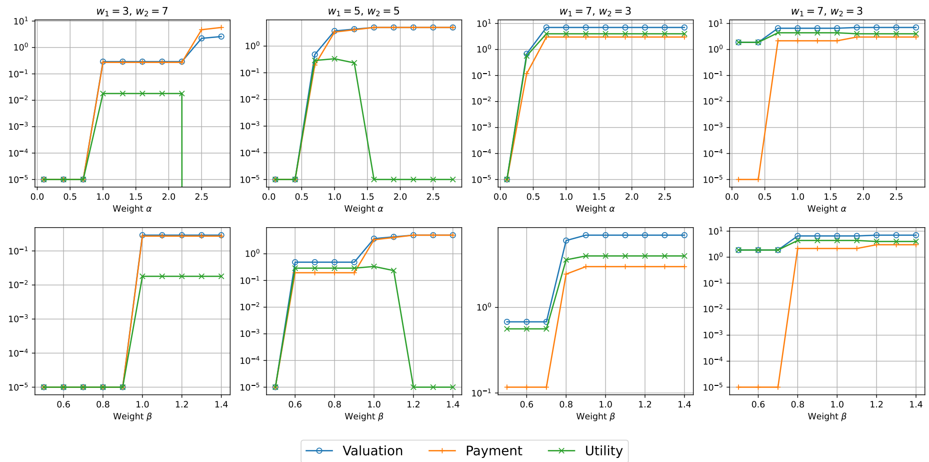

Figure 2: The empirical result for the mechanism (ψ, p^{AFF}). We set the group number $n = 2$, and the group size for each column is in the title. The first three columns are for the "Harmless v.s. Humor" in the Helpful Assistants task, and the last column is for the "Faithful v.s. Summary" in the Reddit Summary task. We report the valuation, the payment, and the utility for group 1 for different reporting parameters α and β (defined in Appendix C). As is shown in the figure, truthfully report, i.e., $\alpha = 1$ and $\beta = 1$, brings the highest utility for all cases, showcasing the DSIC property of the mechanism.

Additionally, there is a line of work studying other LLMs-related scenarios from the algorithmic game theory perspective. Laufer et al. [2023] abstracts the fine-tuning process as a bargaining game and characterizes the perfect sub-game equilibria. Dubey et al. [2024] proposes an auction where bidders compete to place their content within a summary generated by an LLM. Conitzer et al. [2024] considers incorporating social choice theory in LLM alignment. Feizi et al. [2023] explores the potential for leveraging LLMs in online advertising systems.

Our work is also related to classic studies on auction design (Myerson [1979, 1981], Nisan and Ronen [1999]) and facility location problems (Owen and Daskin [1998], Drezner and Hamacher [2004]). In facility locations, agents can benefit by misreporting a more polarized preference. The idea of such a strategy is similar to our model. However, the reporting strategies can be more complex 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. Further, combined with the agents' discretized input spaces, most of our results cannot be directly derived from existing literature.

D.2 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 (Ramé et al. [2024], Wu et al. [2024], Jang et al. [2023], Chakraborty et al. [2024], Shi et al. [2024], Yang et al. [2024], Rame et al. [2024]). They develop more efficient algorithms to extend the Pareto frontier among different objectives (Rame et al. [2024], Jang et al. [2023], Shi et al. [2024], Yang et al. [2024]) and balance issues from various perspectives (Park et al. [2024], Chakraborty et al. [2024], Ramé et al. [2024]).

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 (Coste et al. [2023], Zhang et al. [2024a]), mitigate the influence of incorrect and ambiguous preferences in the dataset (Wang et al. [2024]), and reduce reward hacking (Eisenstein et al. [2023]). 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.

D.3 Multi-parameter Auctions

Several studies have explored the properties relevant to our paper in various multi-parameter auction scenarios, such as implementability (Rochet [1987], Miyake [1998], Conitzer and Sandholm [2004], Saks and Yu [2005], Bikhchandani et al. [2006], Ashlagi et al. [2010]) and payment equivalence (Ivanova-Stenzel and Salmon [2008], Heydenreich et al. [2009], Bergemann and Välimäki [2010], Pavan et al. [2014]). 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 [1981], 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 (Sandholm and Likhodedov [2015], Roberts [1979], Likhodedov and Sandholm [2004], Briest et al. [2010], Tang and Sandholm [2012], Jehiel et al. [2007]), and proposes optimizations to enhance empirical performance (Curry et al. [2022], Duan et al. [2024a,b]). 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.

D.4 Game Theory and LLMs

In addition to the work we review in the primarily related work, there are others that explored the intersection of game theory and large language models from different perspectives. Some research has proposed algorithms for training LLMs inspired by concepts in game theory, such as Nash learning from human feedback (Munos et al. [2023]), consensus game (Jacob et al. [2023]), and direct Nash optimization (Rosset et al. [2024]), and Gemp et al. [2024].

Furthermore, various studies assess LLMs from a game-theoretical perspective, examining aspects such as rationality (Chen et al. [2023], Fan et al. [2023]), behavior in matrix games (Akata et al. [2023], Gandhi et al. [2023], Lorè and Heydari [2023]), and performance in strategic games like auctions (Guo et al. [2023, 2024a]), Werewolf (Xu et al. [2023a,b]), Avalon (Wang et al. [2023]), Diplomacy (Mukobi et al. [2023], [FAIR]), card game (Feng et al. [2024]) and electronic game (Agashe et al. [2023], Ma et al. [2023], Shao et al. [2024]). There are also comprehensive surveys (Zhang et al. [2024b], Gallotta et al. [2024], Guo et al. [2024b]).

E Omitted proofs in Section 3

Theorem 3.2. *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.*

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 $\mu \in \mathbb{R}$ (Luenberger et al. [1984]), 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 the definition of valuation function, $\frac{\partial v_i}{\partial \text{LLM}_\theta(\boldsymbol{x})} = \text{rm}_i(\boldsymbol{x})$, so we have

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

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

The Summation Normalization Case.

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

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

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 $\theta = \psi((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), \theta_{\text{init}})$ and $\theta' = \psi((\text{rm}'_i, \overrightarrow{\text{rm}}_{-i})), \theta_{\text{init}})$, we use μ and μ' to denote the variable in the necessary condition for LLM_θ and $\text{LLM}_{\theta'}$, and we can derive the following results.

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

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

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

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

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

(b) The order of $\text{LLM}_\theta(\boldsymbol{x})$ and $\text{LLM}_{\theta'}(\boldsymbol{x})$ 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 $\text{LLM}_{\theta'}(\boldsymbol{x}_3) \geq \text{LLM}_\theta(\boldsymbol{x}_3)$. Then we have

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

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

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

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

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

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

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

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

Note that both (2) and (3) are because of $\text{rm}_i(\boldsymbol{x}_1) \geq \text{rm}_i(\boldsymbol{x}_2)$. And unless $\text{rm}_i(\boldsymbol{x}_1) = \text{rm}_i(\boldsymbol{x}_2)$, 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 maximum is similar. We take the $\boldsymbol{x}_1 \in \arg\min_{\boldsymbol{x} \in T^*} \text{rm}_i(\boldsymbol{x})$. Since $\min_{\boldsymbol{x} \in T^*} \text{rm}_i(\boldsymbol{x}) > 0$, we have $\text{rm}_i(\boldsymbol{x}_1) > 0$. Then we take a small $\epsilon < \text{rm}_i(\boldsymbol{x}_1)$ and define rm'_i as:

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

With the same technique, we first show that $\text{LLM}_{\theta'}(\boldsymbol{x}_1) < \text{LLM}_{\theta}(\boldsymbol{x}_1)$ and $\text{LLM}_{\theta'}(\boldsymbol{x}) > \text{LLM}_{\theta}(\boldsymbol{x})$ 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

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

□

Lemma E.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 Condition 3.5.*

Proof. As is shown in the proof of Theorem 3.2, we have two Lagrangian variables μ and μ' for $(\vec{\text{rm}}, \vec{w})$ and $(\vec{\text{rm}}, \vec{w})$, respectively. We have the following equations:

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

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

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

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

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

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

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

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

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

Theorem E.2 (Detailed version of Theorem A.1)**.** *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 $(\vec{rm}_{-i}, \vec{w}_{-i})$ 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 $\theta = \psi(\vec{rm}, \vec{w}, \theta_{init})$. 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 3.2, the necessary condition for the solution θ is that there exists a $\mu \in \mathbb{R}$ such that

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

And by Lemma E.1, we can also use the condition Condition 3.5.

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 $0 < \epsilon < 1 - \text{rm}_i(\boldsymbol{x}_1)$ to construct a report rm'_i

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

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

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

With a similar analyze in the proof of Theorem 3.2, we can induce that $\mu' > \mu$ and $\text{LLM}_{\theta'}(\boldsymbol{x}) < \text{LLM}_{\theta}(\boldsymbol{x})$ 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 $\text{LLM}_{\theta'}(\boldsymbol{x}) \leq \boldsymbol{z} \leq \text{LLM}_{\theta}(\boldsymbol{x})$ such that

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

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

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

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

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

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 $\max_{\boldsymbol{x}} |\text{LLM}_{\theta}(\boldsymbol{x}) - \text{LLM}_{\theta'}(\boldsymbol{x})| < \delta_2$, we have

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

Besides, because of the Condition 3.5, there exists δ_3, such that for each $(\vec{w}, \overrightarrow{\text{rm}})$ and $(\vec{w}', \overrightarrow{\text{rm}}')$ satisfying $\max_{\boldsymbol{x}\in T^*} |\sum_{i=1}^n w_i \text{rm}_i(\boldsymbol{x}) - \sum_{i=1}^n w'_i \text{rm}'_i(\boldsymbol{x})| \leq \delta_3$, we have $\max_{\boldsymbol{x}} |\text{LLM}_{\theta}(\boldsymbol{x}) - \text{LLM}_{\theta'}(\boldsymbol{x})| < \delta_2$.

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

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

This means that

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

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 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 $\boldsymbol{x}_2$, $\mathrm{rm}_i(\boldsymbol{x}_1) \ge \mathrm{rm}_i(\boldsymbol{x}_2) > 0$, 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, $\mathrm{LLM}_{\theta'}(\boldsymbol{x}_2) \le \mathrm{LLM}_{\theta''}(\boldsymbol{x}_2) \le \mathrm{LLM}_\theta(\boldsymbol{x}_2)$ and $\mathrm{LLM}_\theta(\boldsymbol{x}_1) \le \mathrm{LLM}_{\theta''}(\boldsymbol{x}_1) \le \mathrm{LLM}_{\theta'}(\boldsymbol{x}_1)$ such that

$$\begin{aligned}
\mu' - \mu &= w_i\epsilon - f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x}_1)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1)})\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x}_1) - \mathrm{LLM}_\theta(\boldsymbol{x}_1)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1)}, \\
\mu' - \mu &= -w_i\epsilon - f''(\frac{\mathrm{LLM}_{\theta''}(\boldsymbol{x}_2)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)})\frac{\mathrm{LLM}_{\theta'}(\boldsymbol{x}_2) - \mathrm{LLM}_\theta(\boldsymbol{x}_2)}{\mathrm{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)}.
\end{aligned} \tag{3}$$

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

$$\max_{\boldsymbol{x} \in T^*} \left| f''(\frac{\text{LLM}_\theta(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) - f''(\frac{\text{LLM}_{\theta'}(\boldsymbol{x})}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x})}) \right| \le \frac{\frac{\delta_1}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)} f''\left(\frac{\text{LLM}_\theta(\boldsymbol{x}_2)}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_2)}\right) - \frac{\delta_1}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1)} f''\left(\frac{\text{LLM}_\theta(\boldsymbol{x}_1)}{\text{LLM}_{\theta_{\text{init}}}(\boldsymbol{x}_1)}\right)}{3}. \tag{4}$$

We take $\epsilon < \frac{\delta_2}{w_i}$ and first prove that when taking such ϵ, there is $\mu \le \mu'$. By contradiction, if $\mu' < \mu$, then by condition Equation (OPT), for all $\boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$, there is $\text{LLM}_{\theta'}(\boldsymbol{x}) > \text{LLM}_\theta(\boldsymbol{x})$. Therefore, $\text{LLM}_{\theta'}(\boldsymbol{x}_1) - \text{LLM}_\theta(\boldsymbol{x}_1) = \sum_{\boldsymbol{x} \notin \{\boldsymbol{x}_1, \boldsymbol{x}_2\}} (\text{LLM}_\theta(\boldsymbol{x}) - \text{LLM}_{\theta'}(\boldsymbol{x})) + \text{LLM}_\theta(\boldsymbol{x}_2) - \text{LLM}_{\theta'}(\boldsymbol{x}_2) < \text{LLM}_\theta(\boldsymbol{x}_2) - \text{LLM}_{\theta'}(\boldsymbol{x}_2)$. However, by Equation (3), if $\mu' < \mu$, we get

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

By Equation (4), we can derive that

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

and thus, we get

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

, which brings the contradiction.

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

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

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:

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

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. □

Theorem 3.3. *For any $\psi \in \Psi^{SW}$, mechanism (ψ, p^{AFF}) satisfies DSIC and IR.*

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

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

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

$$v_i = w_i v_i(\theta; \text{rm}_i),$$
$$v_i' = w_i v_i(\theta'; \text{rm}_i).$$

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

$$p_i = ASW_{-i}(\psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}); \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) - ASW_{-i}(\theta; \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})$$
$$p_i' = ASW_{-i}(\psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}); \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}}) - ASW_{-i}(\theta'; \overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})$$

Therefore, we can calculate the change in the utility:

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

The last inequality holds by the definition of θ

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

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

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

(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:

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

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

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

□

Proposition E.3. *Condition 3.5 holds for 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, we need λ is relatively large for f_2). For $f_{\text{KL}}(x)$, $\delta = \min\{\frac{\lambda}{2} \log \frac{1}{1-\epsilon}, \frac{\lambda}{2} \log(1+\epsilon)\}$. For $f_2(x)$, $\delta = \lambda\epsilon$.*

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:

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

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

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

Note that we also have the condition

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

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

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

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 $\text{LLM}_\theta(\boldsymbol{x})$ and $\text{LLM}_{\theta'}(\boldsymbol{x})$ can be bounded by

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

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

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

(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:

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

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

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

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

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

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

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

□

Theorem 3.6. *An implementable training rule ψ satisfies payment equivalence if Condition 3.5 holds and $\forall i$, $\overrightarrow{rm}_{-i}$, $\vec{w}_{-i}$, 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 $w_i \in \mathcal{W}$. For maximum normalization, that 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

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

We prove the following lemma, which is a special case in Heydenreich et al. [2009],

Lemma E.4 (Heydenreich et al. [2009]). *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) satisfies DSIC. For any two types t_i, t_i' and a finite and distinct sequence $[t_i', t_i^1, \cdots, t_i^k, t_i]$, let $t_i^0 = t_i'$ and $t_i^{k+1} = t_i$, we have that

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

This can be rewritten as

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

Sum over j, we get the following inequality

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

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 (6) 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 3.6 is sufficient for the condition stated in Lemma E.4. 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 (Rochet [1987]). 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 $[t_i, t_i^1, \cdots, t_i^k, t_i']$ such that

$$\sum_{j=0}^{k} l(t_i^j, t_i^{j+1}) + \sum_{j=0}^{k} l(t_i^{j+1}, t_i^j) \leq \epsilon. \tag{5}$$

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 (5) 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 given by the condition in Theorem 3.6. 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 to 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 E.4, this is a sufficient condition for the payment equivalence of ψ. □

Corollary 3.7. *When Condition 3.5 holds, each 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 3.7 when the mechanism uses SW-Max training rules.

$$\text{rm}^*(x)=\begin{cases}\dfrac{1}{|T^*|} & \text{Summation Normalization Case,}\\ 1 & \text{Maximum Normalization Case.}\end{cases}\tag{6}$$

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

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

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

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

or

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

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

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

or

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

Both cases contradicted the optimality of θ'. □

F Omitted proofs in Section B

Lemma F.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*

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

Proof. We can derive that

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

□

Lemma B.2. *When Condition B.1 holds and the training rule* $\psi \in \Psi^{SW}$*, if all groups truthfully report, the loss in social welfare is bounded by* $2\epsilon \sum_{i=1}^n w_i$*.*

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.

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

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

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

□

Theorem B.3. *When Condition B.1 holds and the training rule $\psi \in \Psi^{SW}$, for group i and any rm_i, rm'_i, $\overrightarrow{rm}_{-i}$, w_i and $\vec{w}_i$, we have*

$$U_i((rm_i, \overrightarrow{rm}_{-i}), (w_i, \vec{w}_{-i}); \psi, p^{AFF}, rm_i, w_i) \geq \\ U_i((rm'_i, \overrightarrow{rm}_{-i}), (w_i, \vec{w}_{-i}); \psi, p^{AFF}, rm_i, w_i) - 2w_i\epsilon.$$

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

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

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

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

where we define $\theta = \psi((\text{rm}'_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}, \theta_{\text{init}})$, and $\theta_{-i} = \psi(\overrightarrow{\text{rm}}_{-i}, \vec{w}_{-i}, \theta_{\text{init}})$. Note that the term $ASW_{-i}(\theta_{-i}; \overrightarrow{\text{rm}}, \vec{w}, \theta_{\text{init}})$ 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) + 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[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[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) + 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 F.1). (2) and (5) hold by the definitions: $\hat{\theta} = \psi((\widehat{\text{rm}}_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}, \theta_{\text{init}})$

$= \arg\max_{\theta' \in \Theta} ASW(\theta'; (\widehat{\text{rm}}_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}, \theta_{\text{init}})$ and $\theta = \psi((\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}, \theta_{\text{init}}) = \arg\max_{\theta' \in \Theta} ASW(\theta'; (\text{rm}_i, \overrightarrow{\text{rm}}_{-i}), \vec{w}, \theta_{\text{init}})$. And (4) holds since the inner term is irrelevant to $\widehat{\text{rm}}_i$.

Therefore, we get

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

□