Online Learning from Strategic Human Feedback in LLM Fine-Tuning

Shugang Hao
Pillar of Engineering Systems and Design
Singapore University of Technology and Design
Singapore, Singapore
shugang_hao@sutd.edu.sg

Lingjie Duan
Pillar of Engineering Systems and Design
Singapore University of Technology and Design
Singapore, Singapore
lingjie_duan@sutd.edu.sg

***Abstract*—Reinforcement learning from human feedback (RLHF) has become an essential step in fine-tuning large language models (LLMs) to align them with human preferences. However, human labelers are selfish and have diverse preferences. They may strategically misreport their online feedback to influence the system's aggregation towards their own preferences. Current practice simply averages labelers' feedback per time and fails to identify the most accurate human labeler, leading to linear regret $\mathcal{O}(T)$ for T time slots. To our best knowledge, we are the first to study online learning mechanisms against strategic human labelers in the LLM fine-tuning process. We formulate a new dynamic Bayesian game and dynamically adjust human labelers' weights in the preference aggregation, ensuring their truhtful feedback and sublinear regret $\mathcal{O}(T^{1/2})$. Simulation results demonstrate our mechanism's great advantages over the existing benchmark schemes.**

***Index Terms*—LLM fine-tuning, online learning, strategic human feedback, truthful mechanism design, regret analysis.**

I. Introduction

Large language models (LLMs) such as ChatGPT and SORA have succeeded in handling a number of tasks such as text and video generation. To better meet users' demands for specific applications, pre-trained LLMs are fine-tuned to be customized using task-oriented datasets (e.g., [1]). Traditional supervised learning methods fail to align with human preferences because of the difficulty in acquiring a significant number of question-answer paired data (e.g., [2], [3]). Reinforcement learning from human feedback (RLHF) has emerged as a promising approach to tackle this human preference alignment problem (e.g., [4], [5]). It queries online human feedback (e.g., in a weekly cadence [6], [7], [8]) to obtain a human-annotated preference dataset, which will then be used to train and update the learning policy. RLHF has become an essential training step in LLM fine-tuning due to its effectiveness in aligning with human preferences.

However, human labelers are selfish in the RLHF loop to have diverse preferences and they may strategically misreport their online feedback to influence the system's aggregation towards their own preferences (e.g., [3], [9]). For example, a user in an LLM rating system may strategically give an extreme response rating of 0 or 10 in the range of [0, 10] to maximally influence the overall rating toward his actual rate (e.g., [10]). Besides, there is a renowned "wet bias" where a weather forecaster as human labeler or predictor may deliberately report an exaggerated probability of precipitation to increase the influence of his forecast in the system's final prediction (e.g., [11]). Current practice of LLM fine-tuning largely ignores human labelers' misreporting and simply averages human feedback in the preference aggregation (e.g., [4], [5], [12], [13], [14], [15], [16]) and we wonder its actual performance. Our first question naturally arises:

- *Q1. How bad is the current practice of average feedback aggregation for LLM fine-tuning performance?*

Later we prove that the average feedback aggregation scheme fails to identify the most accurate human labeler in the online learning process and incurs a non-vanishing regret $\mathcal{O}(T)$ overtime. This motivates us to propose new schemes for truthful human feedback and vanishing regret. In the recent RLHF literature, Sun *et al.* (2024) in [3], Park *et al.* (2024) in [17], Soumalias *et al.* (2024) in [9] and Dubey *et al.* (2024) in [18] focus on monetary payment-based mechanism design to reward and enable strategic human labelers' truthful preference feedback. In practice, monetary mechanisms involve complicated billing issues and may not be easy to implement. Furthermore, these works assume a one-shot or offline preference feedback setting and do not consider human labelers' online feedback. In the online setting, human labelers have more room to strategically misreport and play with the RLHF system for long-term influence.

In the related literature of algorithmic game theory, there are relevant non-monetary mechanism studies on facility location games (e.g., [19], [20], [21]), where the system aims to incentivize customers' truthful reporting of their locations to optimize facility placement. There each customer can strategically misreport his location to mislead the facility placement as close to his location (preference) as possible. The popular "median" scheme (e.g., [10], [22]) to aggregate multi-agent reports is widely used to return customers' truthful reporting. Yet, later we prove that it is no longer truthful for our online problem and also incurs a non-vanishing regret as the average feedback aggregation scheme. Our second question is thus:

- *Q2. How to design an efficient and truthful mechanism for online learning from strategic human feedback in LLM fine-tuning?*

A natural non-monetary mechanism approach is to deploy a weighted aggregation and dynamically adjust each human labeler's weight in the online RLHF process. However, it is challenging to motivate human labelers' truthful preference feedback for a vanishing regret. Human labelers' preferences are hidden and can vary over time slots, which makes it difficult for the system to verify and correct their misreports to learn their preferences (e.g., [3]). Since the most accurate human labeler is unknown in the online learning process, it is difficult for the system to dynamically weigh each human labeler to achieve a vanishing regret.

To our best knowledge, we are the first to study online learning mechanisms against strategic human labelers in LLM fine-tuning. We formulate a new dynamic Bayesian game and dynamically adjust each human labeler's weight in the aggregation according to their feedback accuracy, which ensures their truthful feedback and sublinear regret $\mathcal{O}(T^{1/2})$. Finally, simulation results also demonstrate our mechanism's great advantages over the existing benchmark schemes.

II. System Model and Problem Formulation

First, we introduce our system model. Then, we formulate a new dynamic Bayesian game and give desired properties for guiding our late mechanism design.

A. System Model of LLM Fine-Tuning from Online Human Feedback

We consider an LLM fine-tuning process of aggregating online feedback from $N\geq 2$ strategic human labelers overtime. It starts from fine-tuning a pre-trained language model with supervised learning based on a task-orientated dataset (e.g., paragraph summary) to obtain a reference policy $\pi_{\texttt{ref}}$. Then, the system iterates the RLHF process in T time slots (e.g., a weekly cadence [6], [7], [8]), where each time slot $t \in [T] := \{1, \cdots, T\}$ contains the following three stages. The current practice of LLM fine-tuning only involves Stages I and II with uniform weights in the aggregation and we introduce Stage III to dynamically adjust human labelers' weights according to their online feedback accuracy.

Stage I. Online Preference Feedback: The system draws m_t prompts $\{x_j^t\}_{j=1}^{m_t}$ from the context space $\mathcal{X}$ and generates m_t pairwise responses $\{(y_{l_j}^t, y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}$ from the response space $\mathcal{Y}$ according to the last slot policy π_{t-1} with $\pi_0 = \pi_{\text{ref}}$ (e.g., [8]). It then shares $\{(x_j^t, y_{l_j}^t, y_{l'_j}^t)\}_{j=1}^{m_t}$ with N human labelers for their preference feedback. Each human labeler $i\in[N]$ independently realizes his continuous private preference of response $y_{l_j}^t$ over $y_{l'_j}^t$ as $\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) \in [0,1]$ for each $j \in [m_t]$ and he believes that the ground-truth preference $p_j^t \sim Bernoulli(\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t))$, where realization $p_j^t = 1$ means response $y_{l_j}^t$ is preferred over $y_{l'_j}^t$ and $p_j^t = 0$ otherwise.

He wants to influence the system's aggregation toward his own preference and may feedback another continuous $\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) \in [0,1]$ different from his actual $\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)$ to the system (e.g., [3], [9]). The system and other human labelers are uncertain of his $\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)$ realization.

Stage II. Online Feedback Aggregation and Policy Optimization: After receiving each human labeler i's feedback $\{\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}$ for $i \in [N]$, the system aggregates according to the weight w_i^t for each prompt $j \in [m_t]$ as

$$\mathcal{P}(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) = \frac{\sum_{i=1}^N w_i^t \hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)}{\sum_{i'=1}^N w_{i'}^t}, \quad (1)$$

with a uniform weight w_i^1=1 for all $i\in[N]$ in the first time slot.[1] The aggregated preference $\{\mathcal{P}(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}$ will be included to construct the human preference dataset $\mathcal{D}_t := \{\mathcal{P}(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}$. Based on the preference dataset $\mathcal{D}_t$, the system then learns a policy π_t using direct preference optimization (DPO) to solve a KL-regularized optimization problem against the reference policy π_{ref} (e.g., [12]):

$$\min_{\pi_t} -\mathbb{E}_{(x,y,y')\sim\mathcal{D}_t} \ln \sigma\left(\beta \ln \frac{\pi_t(y|x)}{\pi_{\texttt{ref}}(y|x)} - \beta \ln \frac{\pi_t(y'|x)}{\pi_{\texttt{ref}}(y'|x)}\right),$$

where $\sigma(\cdot)$ denotes the logistic function and β is a parameter of evaluating the deviation from the reference policy $\pi_{\texttt{ref}}$.

Stage III. Reweighing Human Labelers: In the RLHF literature, the current practice of LLM fine-tuning simply averages human labelers' feedback in the preference aggregation using uniform w_i^t=1 for all $i\in[N]$ and $t\in[T]$ in (1) (e.g., [4], [5], [12]). To our best knowledge, we are the first to consider dynamic adjustment of each human labeler's weight in the online RLHF process, where the system determines each

$$w_i^{t+1} = f_i(\{\{\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}\}_{i=1}^N, \{p_j^t\}_{j=1}^{m_t}) \quad (2)$$

for the next slot $t+1$'s aggregation according to feedback $\{\{\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}\}_{i=1}^N$ and the realized preference $\{p_j^t\}_{j=1}^{m_t}$. The system implements and tests the obtained policy π_t for customers' practical usage and learns the realized binary preference $p_j^t \in \{1,0\}$ for each prompt $j \in [m_t]$ according to its customers' realized experience, where $p_j^t = 1$ if $y_{l_j}^t$ is preferred than $y_{l'_j}^t$ and 0 otherwise (e.g., [23], [24], [25]).

Each selfish human labeler wants to use $\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)$ to influence the system's aggregation towards his own preference (e.g., [3], [9]). Thus, he wants to obtain a large weight in the system's feedback aggregation in (1) in each time slot $t \in [T]$ and aims to maximize his long-term influence benefit as the cumulative weight over the whole T time slots as follows:

$$u_i(\{\{\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}\}_{t=1}^T) \quad (3)$$
$$= \sum_{t=1}^T w_i^t(\{\{\hat{\mathcal{P}}_i(y_{w_j}^{t-1} \succ y_{l'_j}^t | x_j^{t-1})\}_{j=1}^{m_{t-1}}\}_{i=1}^N, \{p_j^{t-1}\}_{j=1}^{m_{t-1}}).$$

On the other hand, the system wants to improve the feedback accuracy in the aggregation by assigning the largest weight to the most accurate human feedback. However, the best human labeler is unknown in the online iteration. It then turns to reducing the regret between online weighted aggregation and offline choice of the best human labeler in

[1] One can also use w_i^1=1/N without any change for the aggregation result.

hindsight (e.g., [11], [26], [27]), where the performance loss or online regret is defined as the mean square error (MSE) between the system's weighted aggregation in (1) and the realized binary preference as follows:

$$R(T) := \sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\sum_{i=1}^{N} \frac{w_i^t \hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)}{\sum_{i'=1}^{N} w_{i'}^t} - p_j^t \right)^2 - \min_{i \in [N]} \sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t \right)^2. \quad (4)$$

B. Dynamic Bayesian Game Formulation

Based on our system model above, we formulate the multi-agent online learning as a new dynamic Bayesian game:

- In Stage I of each time slot $t \in [T]$, each human labler i with his private preference $\{\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}$ determines his preference feedback $\{\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}$ to maximize his accumulative weight in (3).
- In Stage III of each time slot $t \in [T]$, the system updates each human labeler's weight $w_i^{t+1} = f_i(\{\{\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}\}_{i=1}^{N}, \{p_j^t\}_{j=1}^{m_t})$ for reducing regret in (4).

Note that there is no strategic decision for human labelers or the system in Stage II. We need to carefully design an online aggregation mechanism for ensuring each human labeler's truthful preference feedback and a vanishing regret in time. We define the desired properties as below.

Definition 1 (Truthfulness for Human Feedback): An online weighted aggregation mechanism $\mathcal{M}$ is truthful if each human labeler $i \in [N]$ obtains a larger long-term influence in (3) over the whole T time slots though truthful preference feedback instead of misreporting in the mean time, i.e.,

$$\mathbb{E}\Bigg[\sum_{t=2}^{T-1} w_i^{t+1} \big(\{\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}, \{p_j^t\}_{j=1}^{m_t}, \{\{\hat{\mathcal{P}}_k(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}\}_{k=1,k\neq i}^{N} \big) \Bigg]$$
$$\geq \mathbb{E}\Bigg[\sum_{t=2}^{T-1} w_i^{t+1} \big(\{\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}, \{p_j^t\}_{j=1}^{m_t}, \{\{\hat{\mathcal{P}}_k(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}\}_{k=1,k\neq i}^{N} \big) \Bigg].$$

Definition 2 (High Efficiency in Sublinear Regret $R(T)$ in (4)): An online weighted aggregation mechanism $\mathcal{M}$ is efficient if its time-average regret $R_{\mathcal{M}}(T)/T$ is vanishing in the time slot number T, i.e., $\lim_{T\to\infty} \frac{R_{\mathcal{M}}(T)}{T} = 0$.

III. Two Benchmark Schemes: Average Feedback Aggregation and Median Aggregation

In this section, we analyze two common schemes used in the literature of both RLHF and algorithmic game theory, serving as two fair benchmarks for our mechanism to compare later.

A. Benchmark 1: Average Feedback Aggregation

The current practice of LLM fine-tuning simply averages human labelers' feedback in the preference aggregation using uniform $w_i^t = 1$ for all $i \in [N]$ and $t \in [T]$ (e.g., [4], [5], [12]). Unfortunately, such an average feedback aggregation scheme can lead to a non-vanishing regret as shown below.

Lemma 1: The system's regret in (4) under the benchmark 1 of average aggregation is $R_1(T) = \mathcal{O}(T)$, leading to a non-vanishing time-average regret $\lim_{T\to\infty} \frac{R_1(T)}{T} > 0$.

Benchmark 1 fails to dynamically adjust human labelers' aggregation weights according to their online feedback accuracy. Thus, the most accurate human labeler cannot receive the largest weight in the online learning process, leading to a non-vanishing time-average regret even if $T \to \infty$. We are motivated to find other schemes to reduce the system's regret.

B. Benchmark 2: Median Aggregation Scheme

In the algorithmic game theory literature, the popular "median" scheme is widely used to motivate strategic agents' truthful reporting (e.g., [10], [22]). We define it below.

Definition 3 (Median Aggregation Scheme): The system first re-organizes human labelers' preference feedback $\{\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{i=1}^{N}$ in an increasing order as $\hat{\mathcal{P}}_{k_1,j}^t \leq \cdots \leq \hat{\mathcal{P}}_{k_N,j}^t$ for each prompt $j \in [m_t]$ in each time slot $t \in [T]$. It then chooses the median $\hat{\mathcal{P}}_{k_s,j}^t$ as its preference aggregation, where the index $s = N/2$ if N is even and $s = (N+1)/2$ otherwise.

Since the benchmark 2 independently commits to the median feedback for each prompt in each time slot, at the equilibrium, all the human labelers' feedback will converge to one common point for an equal probability to be the median. We summarize the equilibrium in the following.

Proposition 1: At the equilibrium of the benchmark 2, all the human labelers' feedback $\hat{\mathcal{P}}_i^*(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)$ will converge to an arbitrary point $\hat{\mathcal{P}}(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) \in [0,1]$ for $j \in [m_t]$, $i \in [N]$ and $t \in [T]$, which may not be the same as their own preference.

The median scheme is not truthful since a human labeler may be committed with a positive probability by misreporting than no probability by truthful feedback. Further, it leads to a non-vanishing time-average regret in T in the following.

Lemma 2: The system's regret in (4) under the untruthful median scheme is $R_2(T) = \mathcal{O}(T)$, leading to a non-vanishing time-average regret $\lim_{T\to\infty} \frac{R_2(T)}{T} > 0$.

The median feedback can lead to a total aggregation loss of $\mathcal{O}(T)$ over the T time slots. Yet, there can exist a human labeler holding $\mathcal{P}_k(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) = p_j^t$, incurring zero aggregation loss of the best human labeler in hindsight. The average regret is then non-vanishing even if the time slot number $T \to \infty$. Given non-vanishing regret of both benchmark schemes, we are well motivated to develop a truthful mechanism to substantially reduce the system's regret.

IV. Online Weighted Aggregation Mechanism: New Design, Analysis, and Simulations

In this section, we first present our mechanism design and its regret bound. We then run simulations for verification.

A. Mechanism Design and Theoretical Analysis

Unlike benchmark 1, in Stage III of each time slot, we dynamically adjust each human labeler's weight based on his feedback accuracy and assign a larger weight if his feedback is closer to the realized binary preference. We need to carefully design the online mechanism in (2) to ensure that each obtains the largest long-term influence in Definition 1 only with truthful feedback. We define our mechanism as below.

Definition 4 (Online Weighted Aggregation Mechanism): At Stage III of each time slot $t \in [T-1]$, the system updates each human labeler's weight w_i^{t+1} in (2) based on his feedback $\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)$ and the realized binary preference p_j^t:

$$w_i^{t+1} = w_i^t \cdot \left(1 - \alpha \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t\right)^2\right), \quad (5)$$

where $\alpha > 0$ is the step-size parameter.

Intuitively, our mechanism determines each human labeler's weight in time slot $t+1$ based on his feedback accuracy in the previous time slot t. If the squared difference between his feedback and the realized binary preference $(\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t)^2$ is small, his weight w_i^{t+1} will be only reduced by a small value from w_i^t. Though all human labelers' weights are decreasing over time, we care about the relative weighted aggregation as in (1).

Since each human labeler holds a Bernoulli belief on p_j^t, our mechanism satisfies the truthful property as shown below.

Proposition 2: Our mechanism in Definition 4 is truthful, i.e., $\hat{\mathcal{P}}_i^*(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) = \mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)$ for any prompt $j \in [m_t]$, human labeler $i \in [N]$ and time slot $t \in [T]$.

Further, our mechanism is efficient and incurs a vanishing time-average regret in T in the following.

Theorem 1: Our mechanism in Definition 4 incurs the sublinear regret $R_{\mathcal{M}}(T)$=$\mathcal{O}(T^{\frac{1}{2}})$ by choosing step-size $\alpha = \frac{2}{3}\sqrt{\frac{2\ln N}{T}}$, leading to zero time-average regret with $\lim_{T\to\infty} \frac{R_{\mathcal{M}}(T)}{T} = 0$.

According to Theorem 1, our mechanism obviously improves from benchmarks 1 and 2 by distinguishing the best human labeler in the online process as $T \to \infty$. As N increases, the system may find a more accurate human labeler in hindsight. Thus, it chooses a larger step-size α to punish inaccurate human labelers more in the weighted aggregation to retire them. As T increases, the system is more patient to choose a smaller α for selecting the best human labeler in hindsight with more time slots and samples.

B. Simulation Results for Verification

In this subsection, we run simulations to show our mechanism's great improvement from the two benchmark schemes.

Based on human-written form post summaries with TRLX framework (e.g., [28]) for the RLHF process, we consider a task of paragraph summarization (e.g., [12]) and use a supervised fine-tuning model to obtain the reference policy π_{ref}. We use the Reddit TL;DR summarization dataset in [29] to draw each prompt x_j^t as a form post from Reddit, and

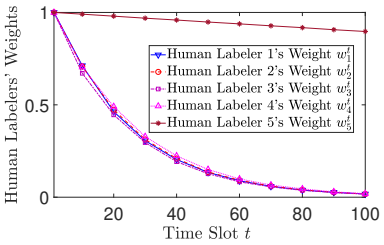

Fig. 1. Each human labeler's weight w_i^t versus time slot t. Here we fix $N = 5$, $T = 100$, $m_t = 50$.

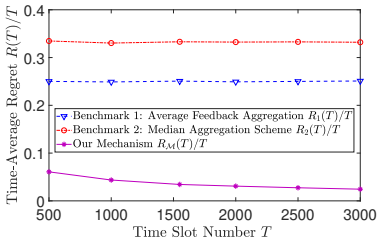

Fig. 2. Time-average regrets of benchmarks 1, 2, and our mechanism versus the time slot number T, respectively. Here we fix $N = 100$ and $m_t = 50$.

the policy π_t generates each pairwise summary $(y_{l_j}^t, y_{l'_j}^t)$ of the main points in the post. We follow [12] to use DPO for the policy training. We use synthetic data to randomly generate each human labeler i's preference $\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)$ in the range of [0, 1] for each pairwise response. Further, we randomly generate the binary realized ground-truth preference $p_j^t \in \{1, 0\}$ for each prompt $j \in [m_t]$, where we fix m_t=50, change N and T for evaluation.

Figure 1 shows how each human labeler's weight w_i^t evolves over time slot $t \in [T]$. We find that our mechanism in Definition 4 dynamically allocates the largest relative weight to the most accurate human labeler 5. As t increases, those inaccurate human labelers 1-4' weights decrease substantially towards 0 to play no effect in RLHF in the end. The system can thus well approximate the real preferences in the offline optimum.

Figure 2 shows the time-average regrets $R(T)/T$ of benchmarks 1, 2, and our mechanism versus the time slot number T. We find that the system's time-average regret is greatly reduced by our mechanism from the two benchmarks. Besides, time-average regrets of both benchmarks 1 and 2 do not decrease with T and are always great than zero, respectively. In constrast, our mechanism's time-average regret decreases with T to approach 0, consistent with Lemmas 1, 2 and Theorem 1.

V. Conclusion

In this paper, we are the first to study online learning from strategic human feedback in LLM fine-tuning. We design an efficient truthful mechanism to achieve zero time-average regret, greatly improving from non-vanishing regrets of the average feedback aggregation and median schemes, respectively. Finally, simulation results demonstrate our mechanism's great advantages over the two benchmark schemes.

References

[1] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Open and efficient foundation language models," *Preprint at arXiv. https://doi. org/10.48550/arXiv*, vol. 2302, 2023.

[2] A. Köpf, Y. Kilcher, D. von Rütte, S. Anagnostidis, Z. R. Tam, K. Stevens, A. Barhoum, D. Nguyen, O. Stanley, R. Nagyfi *et al.*, "Openassistant conversations-democratizing large language model alignment," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[3] H. Sun, Y. Chen, S. Wang, W. Chen, and X. Deng, "Mechanism design for llm fine-tuning with multiple reward models," *arXiv preprint arXiv:2405.16276*, 2024.

[4] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[6] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback. corr, abs/2204.05862, 2022a. doi: 10.48550," *arXiv preprint arXiv.2204.05862*, 2022.

[7] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[8] W. Xiong, H. Dong, C. Ye, Z. Wang, H. Zhong, H. Ji, N. Jiang, and T. Zhang, "Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint," in *Forty-first International Conference on Machine Learning*, 2024.

[9] E. Soumalias, M. J. Curry, and S. Seuken, "Truthful aggregation of llms with an application to online advertising," *arXiv preprint arXiv:2405.05905*, 2024.

[10] V. Conitzer, R. Freedman, J. Heitzig, W. H. Holliday, B. M. Jacobs, N. Lambert, M. Mossé, E. Pacuit, S. Russell, H. Schoelkopf *et al.*, "Social choice for ai alignment: Dealing with diverse human feedback," *arXiv preprint arXiv:2404.10271*, 2024.

[11] T. Roughgarden and O. Schrijvers, "Online prediction with selfish experts," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[12] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] T. Xie, D. J. Foster, A. Krishnamurthy, C. Rosset, A. Awadallah, and A. Rakhlin, "Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf," *arXiv preprint arXiv:2405.21046*, 2024.

[14] S. Cen, J. Mei, K. Goshvadi, H. Dai, T. Yang, S. Yang, D. Schuurmans, Y. Chi, and B. Dai, "Value-incentivized preference optimization: A unified approach to online and offline rlhf," *arXiv preprint arXiv:2405.19320*, 2024.

[15] S. Zhang, D. Yu, H. Sharma, Z. Yang, S. Wang, H. Hassan, and Z. Wang, "Self-exploring language models: Active preference elicitation for online alignment," *arXiv preprint arXiv:2405.19332*, 2024.

[16] L. Chen, J. Chen, C. Liu, J. Kirchenbauer, D. Soselia, C. Zhu, T. Goldstein, T. Zhou, and H. Huang, "Optune: Efficient online preference tuning," *arXiv preprint arXiv:2406.07657*, 2024.

[17] C. Park, M. Liu, D. Kong, K. Zhang, and A. E. Ozdaglar, "Rlhf from heterogeneous feedback via personalization and preference aggregation," in *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*, 2024.

[18] K. A. Dubey, Z. Feng, R. Kidambi, A. Mehta, and D. Wang, "Auctions with llm summaries," *arXiv preprint arXiv:2404.08126*, 2024.

[19] M. Asadi, A. Bellet, O.-A. Maillard, and M. Tommasi, "Collaborative algorithms for online personalized mean estimation," *Transactions on Machine Learning Research Journal*, 2022.

[20] Y. Chen, J. Zhu, and K. Kandasamy, "Mechanism design for collaborative normal mean estimation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[21] J. Li, M. Li, and H. Chan, "Strategyproof mechanisms for group-fair obnoxious facility location problems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 9, 2024, pp. 9832–9839.

[22] Y. Wang, H. Zhou, and M. Li, "Positive intra-group externalities in facility location," in *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 2024, pp. 1883–1891.

[23] A. Pacchiano, A. Saha, and J. Lee, "Dueling rl: reinforcement learning with trajectory preferences," *arXiv preprint arXiv:2111.04850*, 2021.

[24] X. Chen, H. Zhong, Z. Yang, Z. Wang, and L. Wang, "Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 3773–3793.

[25] H. Zhong, G. Feng, W. Xiong, L. Zhao, D. He, J. Bian, and L. Wang, "Dpo meets ppo: Reinforced token optimization for rlhf," *arXiv preprint arXiv:2404.18922*, 2024.

[26] R. Frongillo, R. Gomez, A. Thilagar, and B. Waggoner, "Efficient competitions and online learning with strategic forecasters," in *Proceedings of the 22nd ACM Conference on Economics and Computation*, 2021, pp. 479–496.

[27] R. Freeman, D. Pennock, C. Podimata, and J. W. Vaughan, "No-regret and incentive-compatible online learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 3270–3279.

[28] A. Havrilla, M. Zhuravinskyi, D. Phung, A. Tiwari, J. Tow, S. Biderman, Q. Anthony, and L. Castricato, "trlx: A framework for large scale reinforcement learning from human feedback," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 8578–8595.

[29] M. Völske, M. Potthast, S. Syed, and B. Stein, "Tl; dr: Mining reddit to learn automatic summarization," in *Proceedings of the Workshop on New Frontiers in Summarization*, 2017, pp. 59–63.

APPENDIX

A. Proof of Lemma 1

We will prove $R_1(T) = \mathcal{O}(T)$ with a possible sequence of human labelers' preferences. In particular, we consider $\mathcal{P}_k(y^t_{w_j} \succ y^t_{l_j}|x^t_j) = p^t_j$ holds for one particular $k \in [N]$ with any $j \in [m_t]$ and $t \in [T]$. For the remaining human labelers, we consider $(\sum_{i=1,i\neq k}^{N} \frac{1}{N}\hat{\mathcal{P}}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - \frac{N-1}{N}p^t_j)^2 = c^t_j$ for each $i \neq k, i \in [N]$, $j \in [m_t]$ and $t \in [T]$, where $c^t_j \in [\frac{1}{2}, 1]$. Accordingly, we have the best-fixed human labeler in hindsight is $i^* = k$, which brings

$$\min_{i\in[N]} \sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j \right)^2 = 0.$$

However, with the system's uniform weight scheme, we have the cumulative aggregation loss over T slots as follows:

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\sum_{i=1}^{N} \frac{w^t_i \hat{\mathcal{P}}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j)}{\sum_{i'=1}^{N} w^t_{i'}} - p^t_j \right)^2 \\
=&\sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\sum_{i=1}^{N} \frac{\hat{\mathcal{P}}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j)}{N} - p^t_j \right)^2 \\
=&\sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\sum_{i=1,i\neq k}^{N} \frac{1}{N}\hat{\mathcal{P}}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - \frac{N-1}{N}p^t_j \right)^2 \\
=&\sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} c^t_j = \mathcal{O}(T),
\end{aligned}$$

where the last equality holds because each $c^t_j \in [\frac{1}{2}, 1]$ and $\sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} c^t_j$ does not vanish as $T \to \infty$. Finally, we have the regret for bechmark 1 of average aggregation as follows:

$$R_1(T) = \sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\sum_{i=1}^{N} \frac{w^t_i \hat{\mathcal{P}}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j)}{\sum_{i'=1}^{N} w^t_{i'}} - p^t_j \right)^2 - \min_{i\in[N]} \sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j \right)^2 = \mathcal{O}(T),$$

which indicates that $\lim_{T\to\infty} \frac{R_1(T)}{T} > 0$. We then finish the proof.

B. Proof of Proposition 1

We want to prove that $\hat{\mathcal{P}}^*_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) = \hat{\mathcal{P}}(y^t_{w_j} \succ y^t_{l_j}|x^t_j) \in [0, 1]$ for $j \in [m_t]$, $i \in [N]$ and $t \in [T]$ is an equilibrium. Note that the system independent determines the aggregation as the median feedback for each prompt $j \in [m_t]$ in each time slot $t \in [T]$, each human labeler $i \in [N]$ aims to maximize each w^t_i to obtain a largest possible accumulative weight overtime. Since the system always commits to the median feedback, given the other $N-1$ human labelers except for i choose $\hat{\mathcal{P}}_k(y^t_{w_j} \succ y^t_{l_j}|x^t_j) = \hat{\mathcal{P}}(y^t_{w_j} \succ y^t_{l_j}|x^t_j)$, $k \neq i$, $k \in [N]$, the human labeler i feedback any $\hat{\mathcal{P}}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) \neq \hat{\mathcal{P}}(y^t_{w_j} \succ y^t_{l_j}|x^t_j)$ will lead to his weight $w^t_i = 0$ because such feedback cannot be the median. Thus, he will feedback consistently as $\hat{\mathcal{P}}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) = \hat{\mathcal{P}}(y^t_{w_j} \succ y^t_{l_j}|x^t_j)$ for an equal chance to be the median and will never deviate from this feedback strategy. We then finish the proof.

C. Proof of Lemma 2

We want to prove $R_2(T) = \mathcal{O}(T)$ with a possible sequence of human labelers' preferences. In particular, we consider $\mathcal{P}_k(y^t_{w_j} \succ y^t_{l_j}|x^t_j) = p^t_j$ holds for one particular $k \in [N]$ with any $j \in [m_t]$ and $t \in [T]$. Further, we consider $(\hat{\mathcal{P}}^t_{j,k_m} - p^t_j)^2 = c^t_j$ for $j \in [m_t]$ and $t \in [T]$, where $\hat{\mathcal{P}}^t_{j,k_m}$ denotes the median of human labelers' feedback $\{\hat{\mathcal{P}}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j)\}_{i=1}^{N}$ and $c^t_j \in [\frac{1}{2}, 1]$. Accordingly, we have the best-fixed human labeler in hindsight is $i^* = k$, which brings

$$\min_{i\in[N]} \sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j \right)^2 = 0.$$

However, with the system's median scheme, we have the cumulative aggregation loss over T slots as follows:

$$\begin{aligned}
&\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}\left(\sum_{i=1}^{N}\frac{w_i^t\hat{\mathcal{P}}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t)}{\sum_{i'=1}^{N}w_{i'}^t}-p_j^t\right)^2\\
=&\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}\left(\hat{\mathcal{P}}_{j,k_m}^t-y_j^t\right)^2\\
=&\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}c_j^t=\mathcal{O}(T),
\end{aligned}$$

where the last equality holds because each $c_j^t\in[\frac{1}{2},1]$ and $\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}c_j^t$ does not vanish as $T\to\infty$. Finally, we have the regret of the median scheme as follows:

$$R_2(T)=\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}\left(\sum_{i=1}^{N}\frac{w_i^t\hat{\mathcal{P}}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t)}{\sum_{i'=1}^{N}w_{i'}^t}-p_j^t\right)^2-\min_{i\in[N]}\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}\left(\mathcal{P}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t)-p_j^t\right)^2=\mathcal{O}(T).$$

We then finish the proof.

D. Proof of Proposition 2

Note that each human labeler believes that $p_j^t\sim\texttt{Bernoulli}(\mathcal{P}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t))$, we have expectation on w_i^{t+1} in (5) over p_j^t is

$$\begin{aligned}
&\mathbb{E}[w_i^{t+1}]\\
=&w_i^t\frac{1}{m_t}\sum_{j=1}^{m_t}\left[1-\alpha\mathcal{P}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t)(\hat{\mathcal{P}}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t)-1)^2-\alpha(1-\mathcal{P}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t))(\hat{\mathcal{P}}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t)-0)^2\right]\\
=&w_i^t\frac{1}{m_t}\sum_{j=1}^{m_t}\left[1-\alpha(\hat{\mathcal{P}}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t)-\mathcal{P}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t))^2-\alpha(\mathcal{P}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t)-\mathcal{P}_i^2(y_{w_j}^t\succ y_{l_j}^t|x_j^t))\right],
\end{aligned}$$

which is maximized at $\hat{\mathcal{P}}_i^*(y_{w_j}^t\succ y_{l_j}^t|x_j^t)=\mathcal{P}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t)$. To obtain the largest possible accumulative weight, each human labeler will truthfully feedback his preference in the first time slot and all the following time slots because any deviation will lead to smaller weights of the next and all the following time slots. We then finish the proof.

E. Proof of Theorem 1

According to Proposition 2, we have $\hat{\mathcal{P}}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t)=\mathcal{P}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t)$ for all $j\in[m_t]$, $i\in[N]$ and $t\in[T]$. To derive a lower-bound on $\ln\frac{\sum_{i=1}^{N}w_i^{T+1}}{\sum_{i=1}^{N}w_i^1}$, we have

$$\begin{aligned}
\ln\frac{\sum_{i=1}^{N}w_i^{T+1}}{\sum_{i=1}^{N}w_i^1}=&\ln\left(\sum_{i=1}^{N}w_i^{T+1}\right)-\ln\left(\sum_{i=1}^{N}w_i^1\right)\\
=&\ln\left(\sum_{i=1}^{N}\prod_{t=1}^{T}(1-\alpha\frac{1}{m_t}\sum_{j=1}^{m_t}(\mathcal{P}_i(y_{w_j}^t\succ y_{l_j}^t|x_j^t)-p_j^t)^2)\right)-\ln N\\
\geq&\ln\left(\prod_{t=1}^{T}(1-\alpha\frac{1}{m_t}\sum_{j=1}^{m_t}(\mathcal{P}_{i^*}(y_{w_j}^t\succ y_{l_j}^t|x_j^t)-p_j^t)^2)\right)-\ln N\\
=&\sum_{t=1}^{T}\ln\left(1-\alpha\frac{1}{m_t}\sum_{j=1}^{m_t}(\mathcal{P}_{i^*}(y_{w_j}^t\succ y_{l_j}^t|x_j^t)-p_j^t)^2\right)-\ln N\\
\geq&-\alpha\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}(\mathcal{P}_{i^*}(y_{w_j}^t\succ y_{l_j}^t|x_j^t)-p_j^t)^2-\alpha^2\sum_{t=1}^{T}\left(\frac{1}{m_t}\sum_{j=1}^{m_t}(\mathcal{P}_{i^*}(y_{w_j}^t\succ y_{l_j}^t|x_j^t)-p_j^t)^2\right)^2-\ln N\\
\geq&-\alpha\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}\left(\mathcal{P}_{i^*}(y_{w_j}^t\succ y_{l_j}^t|x_j^t)-p_j^t\right)^2-\alpha^2T-\ln N,
\end{aligned}\tag{6}$$

where we choose $\alpha < \frac{1}{2}$ and denote i^* as the best human labeler in hindsight. The first and the third inequalities hold due to $0 < \alpha < \frac{1}{2}$ and $0 \leq (\mathcal{P}_{i^*}(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j)^2 \leq 1$ for all $i \in [N]$ and $t \in [T]$. The second inequality holds due to $\ln(1-x) \geq -x - x^2$ for $x \leq \frac{1}{2}$.

To derive an upper-bound on $\ln \frac{\sum_{i=1}^N w_i^{t+1}}{\sum_{i=1}^N w_i^t}$, we have

$$\begin{aligned}
&\ln \frac{\sum_{i=1}^N w_i^{t+1}}{\sum_{i=1}^N w_i^t} \\
=&\ln \left(\frac{\sum_{i=1}^N w_i^t \cdot (1 - \alpha \frac{1}{m_t} \sum_{j=1}^{m_t} (\mathcal{P}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j)^2)}{\sum_{i'=1}^N w_{i'}^t} \right) \\
\leq&\ln \left(\frac{\sum_{i=1}^N w_i^t \cdot e^{-\alpha \frac{1}{m_t} \sum_{j=1}^{m_t} (\mathcal{P}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j)^2}}{\sum_{i'=1}^N w_{i'}^t} \right) \\
\leq&-\alpha \frac{1}{m_t} \sum_{j=1}^{m_t} \frac{\sum_{i=1}^N w_i^t (\mathcal{P}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j)^2}{\sum_{i'=1}^N w_{i'}^t} + \frac{\alpha^2}{8},
\end{aligned} \tag{7}$$

where the first inequality holds due to $1 - \alpha x \leq e^{-\alpha x}$ for $0 \leq x \leq 1$ and $\alpha > 0$, the second due to Hoeffding's lemma: for a random variable $X = -\frac{1}{m_t} \sum_{j=1}^{m_t} (\mathcal{P}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j)^2 \in [-1, 0]$ and $\alpha \in R$, we have

$$\ln(\mathbf{E}[e^{\alpha X}]) \leq \alpha \mathbf{E}[X] + \frac{\alpha^2 (1-0)^2}{8}.$$

According to (7), we have

$$\begin{aligned}
\ln \frac{\sum_{i=1}^N w_i^{T+1}}{\sum_{i=1}^N w_i^1} &= \ln \left(\frac{\sum_{i=1}^N w_i^{T+1}}{\sum_{i=1}^N w_i^t} \frac{\sum_{i=1}^N w_i^t}{\sum_{i=1}^N w_i^{t-1}} \cdot \cdots \cdot \frac{\sum_{i=1}^N w_i^2}{\sum_{i=1}^N w_i^1} \right) = \sum_{t=1}^T \ln \frac{\sum_{i=1}^N w_i^{T+1}}{\sum_{i=1}^N w_i^t} \\
&\leq -\alpha \sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} \frac{\sum_{i=1}^N w_i^t (\mathcal{P}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j)^2}{\sum_{i'=1}^N w_{i'}^t} + \frac{\alpha^2 T}{8}.
\end{aligned} \tag{8}$$

According to (6) and (8), we have

$$\begin{aligned}
&-\alpha \sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_{i^*}(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j \right)^2 - \alpha^2 T - \ln N \\
\leq&-\alpha \sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} \frac{\sum_{i=1}^N w_i^t (\mathcal{P}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j)^2}{\sum_{i'=1}^N w_{i'}^t} + \frac{\alpha^2 T}{8}.
\end{aligned}$$

After re-arranging the above inequalities and dividing α on both sides, we have

$$\sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} \frac{\sum_{i=1}^N w_i^t (\mathcal{P}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j)^2}{\sum_{i'=1}^N w_{i'}^t} - \sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_{i^*}(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j \right)^2 \leq \frac{\ln N}{\alpha} + \frac{9T\alpha}{8}.$$

Choosing $\alpha = \frac{2}{3}\sqrt{\frac{2 \ln N}{T}} < \frac{1}{2}$, we have

$$\sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} \frac{\sum_{i=1}^N w_i^t (\mathcal{P}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j)^2}{\sum_{i'=1}^N w_{i'}^t} - \sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_{i^*}(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j \right)^2 \leq 3\sqrt{\frac{T \ln N}{2}}.$$

Finally, we have the regret $R_{\mathcal{M}}(T)$ satisfying

$$\begin{aligned}
R_{\mathcal{M}}(T) &= \sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\sum_{i=1}^N \frac{w_i^t \hat{\mathcal{P}}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j)}{\sum_{i'=1}^N w_{i'}^t} - p^t_j \right)^2 - \sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_{i^*}(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j \right)^2 \\
&\leq \sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} \frac{\sum_{i=1}^N w_i^t (\mathcal{P}_i(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j)^2}{\sum_{i'=1}^N w_{i'}^t} - \sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_{i^*}(y^t_{w_j} \succ y^t_{l_j}|x^t_j) - p^t_j \right)^2 \\
&\leq 3\sqrt{\frac{T \ln N}{2}} = \mathcal{O}(T^{\frac{1}{2}}),
\end{aligned}$$

where the first inequality holds due to the convexity of the aggregation loss function. We then finish the proof.