Aggregating Human Feedback from Strategic Labelers in RLHF: Non-Monetary Mechanism Design

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

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

Abstract

Reinforcement learning from human feedback (RLHF) has become a fundamental training step in LLM fine-tuning to align with human preferences, where the system trains its policy based on a human-annotated dataset. Yet, labelers have diverse preferences and they may misreport their preference feedback to influence the system's final inference towards their preferences. With no access to any labeler's private preference, it is difficult for the system to verify and correct such misreports to learn the ground truth. Existing works focus on monetary/direct payment to incentivize truthful human feedback according to outcomes, requiring complicated billing. Further, sustaining monetary rewards is against voluntary human feedback and becomes prohibitively expensive. In this paper, we study how to truthfully aggregate human feedback from strategic labelers in RLHF. Despite the absence of payment measures, we design a novel PArtial Information Disclosure (PAID) mechanism to ensure labelers' truthful feedback and minimize the system's performance loss. We prove that PAID's performance loss is at most 1/4 of the current practice of RLHF and 2/7 of the popular weighted median scheme, respectively. We also show that our PAID is robust to inaccurate human belief held by the system to well bound the performance loss. Finally, we run experiments using several real-world LLM datasets to demonstrate our PAID's great advantages over the common benchmarks and show its positive "side-effect" to even align with most strategic labelers' preferences.

CCS Concepts

• **Networks** → **Network economics**; • **Computing methodologies** → **Distributed algorithms**.

Keywords

Reinforcement learning from human feedback, truthful learning, strategic human feedback, mechanism design

ACM Reference Format:
Shugang Hao and Lingjie Duan. 2018. Aggregating Human Feedback from Strategic Labelers in RLHF: Non-Monetary Mechanism Design. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (MobiHoc'25).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

MobiHoc'25, Houston, TX

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Large Language Models (LLMs) have succeeded in handling a number of tasks such as text and video generation (e.g., ChatGPT and SORA). To better serve users' demands for specific applications, pre-trained LLMs are fine-tuned to be customized using task-oriented datasets (e.g., [31]). 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., [19, 30]). Reinforcement learning from human feedback (RLHF) has emerged as a promising approach to tackle this human preference alignment problem (e.g., [7, 23]). It trains the policy based on a human-annotated dataset in the standard reinforcement learning process. Due to its effectiveness, RLHF has become a fundamental training step in LLM fine-tuning.

However, recent studies in RLHF (e.g., [30],[29], [24], [8]) find that labelers have diverse preferences and they may misreport their feedback to influence the system's final decision towards their preferences. [8] gives an example of three human feedback in a scale from 0 to 10 in RLHF. If a human labeler does not like a generated response with a preference of 3 and suspects that the other two labelers may like it with a preference of 6, he may strategically feedback 2 or even 0 to offset others' positive feedback, which can mislead the system's final average evaluation from 5 to 4, closer to his preference 3. Strategic labelers' misreports pose a significant risk to the correctness and fairness of the system's final evaluation (e.g., [39], [36], [5]). Therefore, it is important to aggregate labelers' actual preferences and balance in the system's final evaluation.

In the LLM and RLHF literature, some recent studies focus on designing mechanisms to aggregate multiple labelers' diverse preferences. Duetting *et al.* (2024) [12] propose a mechanism on online advertising to determine labelers' payments according to their bids, without considering labelers' misreporting of their preferences (same for [38], [4],[10]). Sun *et al.* (2024) [30], Soumalias *et al.* (2024) [29] and Dubey *et al.* (2024) [11] focus on payment-based mechanism design to learn labelers' truthful preferences. There are some other related studies on incentivizing labelers' truthfulness in the field of collective revelation (e.g., [13]), strategic machine learning (e.g., [32]), and peer-prediction including Bayesian Truth Serum (e.g., [18], [37]), proper scoring rules (e.g., [33]), prediction markets (e.g., [17]). Yet, all the above studies focus on monetary payment incentives. In practice, monetary mechanisms involve complicated billing issues and may not be easy to implement. Further, sustaining monetary incentives for a large pool of human feedback can become prohibitively expensive, especially for extensive datasets

and long-term projects (e.g., [15]). It is thus desirable to design non-monetary truthful information disclosure mechanisms for human feedback.

There are two challenges to aggregate human labelers' actual preferences and offset for a fair final evaluation. On the LLM system side, with no access to any labeler's private preference, the system finds it difficult to verify and correct their misreports to learn the ground truth (e.g., [30]). On the labeler side, they can strategically manipulate their feedback against the system's learning or inference, making it even harder for the system to learn the ground truth. The current practice of RLHF simply takes an average of strategic labelers' preference feedback (e.g., [23], [26],[28]), neglecting their strategic manipulation for their own benefit. Our first question naturally arises:

- *Q1. How bad is the current practice of RLHF facing strategic human feedback? How do the existing mechanisms perform?*

Later we prove that the current practice of RLHF performs poorly. This motivates us to find some other truthful mechanisms. There are mainly two streams of work on non-monetary/indirect mechanisms in the literature. On one hand, conventional wisdom studies truthful mechanism design for cheap-talk games, where information senders (i.e., labelers) observe the nature state and send messages to proactively affect the receiver's (system's) inference of the actual state (e.g., [14], [22], [2]). Yet, these mechanisms cannot fit into our learning setting. They consider either one or two labelers and assume uniformity in their data, which cannot be used for our scenario with an arbitrary labeler number and diverse labelers' preferences or loss objectives.

On the other hand, in the algorithmic game theory literature, there are relevant studies on facility location games (e.g., [3], [6], [20]), where the system aims to incentivize customers' truthful reporting of their locations to optimize facility placement close to the mean point. There each customer can strategically misreport his location to mislead the facility placement as close to his location (preference) as possible. The popular "weighted median" scheme (e.g., [8], [35]) is widely used to return customers' truthful reporting as a non-monetary/indirect mechanism. Yet, later we prove that it can perform even worse than the RLHF practice. As such, our second question is:

- *Q2. How to design truthful mechanisms to aggregate human feedback from strategic labelers and reduce the system's performance loss as much as possible?*

We summarize our key novelty and main results below.

- *Aggregating human feedback from strategic labelers in RLHF:* In this paper, we focus on how to truthfully aggregate human feedback from strategic labelers in RLHF. Unlike the LLM literature on aggregating multiple labelers' preferences via monetary incentives (e.g., [30], [29],[24], [12],[11]), we study non-monetary/indirect mechanism design for ease of implementation in RLHF practice. We find that the current practice of RLHF performs poorly. We aim for new analytical studies to guide *how a system can best aggregate strategic human feedback for a fair evaluation.*
- *A new truthful partial information disclosure mechanism:* Though the popular weighted median scheme ensures labelers' truthful feedback, we prove that it may perform even worse than the RLHF practice. Therefore, we no longer expect each labeler to feedback precisely a point but a range of preference. We design a novel PArtial Information Disclosure (PAID) mechanism to aggregate truthful preference feedback from strategic labelers while minimizing the final performance loss. We manage to transform the PAID design problem into solving a linear, autonomous, second-order difference equation in closed form for labelers' uniformly distributed preferences. We prove that our PAID's error is at most 1/4 of the RLHF practice and 2/7 of the weighted median.
- *Robustness to labelers' general preference distributions and the system's inaccurate human belief:* To let our PAID mechanism fit into any distribution of labelers' preferences, we propose an efficient algorithm, which is scalable for a large group of labelers and only incurs a quadratic convergence rate. Furthermore, we prove that our PAID is robust to inaccurate human belief held by the system to well bound the performance loss. Finally, we run experiments using several real-world LLM datasets to demonstrate our PAID's great advantages over the common benchmarks and show its positive "side-effect" to even align with most strategic labelers' preferences.

The rest of this paper is organized as follows. Section 2 introduces the system model on reinforcement learning from strategic human feedback and the dynamic Bayesian game formulation. Section 3 analyzes two common schemes used in the literature as benchmarks for our PAID to compare later. Section 4 details our PAID design and analysis. Section 5 compares the performance of our PAID with the two common benchmark schemes and proposes an efficient algorithm for our PAID design for labelers' general preference distributions. Section 6 extends our PAID design for inaccurate human belief held by the system. Section 7 presents experimental results based on several real-world datasets. Section 8 finally concludes.

2 System Model and Problem Formulation

In Section 2.1, we introduce our system model. In Section 2.2, we formulate our dynamic Bayesian game and give desired properties for guiding mechanism design in Section 4.

2.1 System Model of Aggregating from Strategic Human labelers

As shown in Figure 1, we consider an RLHF process of aggregating feedback from $N \geq 2$ strategic labelers in four stages:

- Stage 0: The system draws a prompt x from the context space $\mathcal{X}$, generates pairwise responses $(y, y'|x)$ from the response space $\mathcal{Y}$, and shares the prompt and responses $(y, y'|x)$ with N labelers for their preference feedback.
- Stage I: After receiving $(y, y'|x)$ from the system, each strategic labeler $i \in [N] := \{1, \cdots, N\}$ independently realizes his continuous private preference $\theta_i = p_i(y \succ y'|x)$ as the probability of preferring response y over y' for prompt x, which is within a bounded interval $[A_i, B_i] \subseteq [0, 1]$ (e.g., [29],[21],[26]). Focusing on his own loss function, he expects the system to adopt his preference as the final one and may feedback another $m_i(\theta_i)$ different from his actual θ_i to the

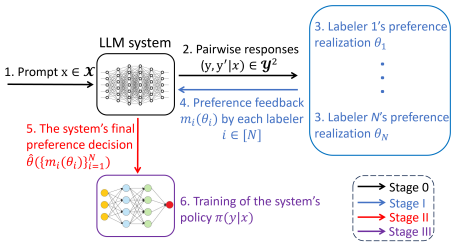

Figure 1: System model illustration in four stages. In Stage 0, the system shares a prompt and its generated pairwise responses with N labelers for preference feedback. In Stage I, N labelers realize their private preferences $\{\theta_i\}_{i=1}^N$ and feedback $\{m_i(\theta_i)\}_{i=1}^N$ to the system, where $m_i(\theta_i)$ may be manipulated to differ from θ_i. In Stage II, the system determines the final evaluation $\hat{\theta}(\{m_i(\theta_i)\}_{i=1}^N)$ based on labelers' feedback to best infer the fair weighted preference $\sum_{i=1}^N w_i\theta_i$, which will be used for training the system's policy in Stage III.

system. The system and the other labelers have no access to his private preference and are uncertain of his θ_i realization. Instead, they can only infer from his historical feedback that θ_i in the range of $[A_i, B_i]$ is roughly distributed according to a PDF $f_i(\cdot)$ with mean μ_i and variance σ_i^2 . Later in Section 6, we extend our mechanism and design to address a more challenging case where the system holds even inaccurate inference or belief of each labeler i's θ_i distribution.

- Stage II: According to N labelers' feedback $\{m_i(\theta_i)\}_{i=1}^N$, the system determines a final preference $\hat{\theta}(\{m_i(\theta_i)\}_{i=1}^N)$ to approximate the weighted-average $\sum_{i=1}^N w_i\theta_i$, where labeler i's weight w_i is pre-determined based on record of his past feedback quality and accuracy with $\sum_{i=1}^N w_i{=}1$ (e.g., [8], [40],[25]). Note that $w_i = \frac{1}{N}$ for $i \in [N]$ is a special case.
- Stage III: The system repeats Stages 0-II by drawing a number of distinct prompts and generating corresponding pairwise responses for human feedback to aggregate. Each aggregated preference $\hat{\theta}(\{m_i(\theta_i)\}_{i=1}^N)$ will be included to construct the human preference dataset $\mathcal{D}$. Based on the preference dataset $\mathcal{D}$, the system then learns a policy π using direct preference optimization (DPO) to solve a KL-regularized optimization problem against the reference policy π_{ref} (e.g., [26]).

Each strategic labeler i only cares about minimizing the square distance between his θ_i preference and the system's final evaluation decision $\hat{\theta}(\{m_i(\theta_i)\}_{i=1}^N)$, yet he is uncertain about the others' preferences θ_{-i} and their feedback strategy $m_{-i}(\cdot)$ from their θ_{-i}. Similar to [9] and [16], we define his loss function $\ell_i(\theta_i, \hat{\theta}(\{m_j(\theta_j)\}_{j=1}^N))$ as the mean square error (MSE) as follows:

$$\ell_i(\theta_i, \hat{\theta}(\{m_j(\theta_j)\}_{j=1}^N)) = \mathbb{E}_{\theta_{-i}, m_{-i}(\theta_{-i})}\left(\theta_i - \hat{\theta}(\{m_j(\theta_j)\}_{j=1}^N)\right)^2. \quad (1)$$

Not favoring a particular labeler's preference, the system wants to fairly obtain the weighted preference $\sum_{i=1}^N w_i\theta_i$ to represent all labelers' preferences, and aims to minimize the MSE between the actual $\sum_{i=1}^N w_i\theta_i$ and its final evaluation $\hat{\theta}(\{m_i(\theta_i)\}_{i=1}^N)$ as follows:

$$L\left(\hat{\theta}(\{m_i(\theta_i)\}_{i=1}^N)\right) = \mathbb{E}_{\{\theta_i\}_{i=1}^N, \{m_i(\theta_i)\}_{i=1}^N}\left(\sum_{i=1}^N w_i\theta_i - \hat{\theta}(\{m_i(\theta_i)\}_{i=1}^N)\right)^2. \quad (2)$$

2.2 Dynamic Bayesian Game Formulation

Based on our system model above, we formulate a dynamic Bayesian game to include each labeler i's feedback $m_i(\theta_i)$ for minimizing (1) in Stage I and the system's inference for final evaluation $\hat{\theta}(\{m_i(\theta_i)\}_{i=1}^N)$ to minimize (2) under incomplete information in Stage II.[1] Note that the system has no access to any labeler i's private preference and cannot verify if a strategic labeler's feedback is true ($m_i(\theta_i){=}\theta_i$) or not. Further, labelers are strategic in manipulating their feedback against the system's inference for their own loss minimization, which makes it even harder to learn their preferences. For example, labeler i expecting $\theta_i{<}\hat{\theta}(\{m_i(\theta_i)\}_{i=1}^N)$ will under-report his preference $m_i(\theta_i){<}\theta_i$ to mislead the final $\hat{\theta}(\{m_i(\theta_i)\}_{i=1}^N)$ closer to θ_i.

To handle the above two challenges, we need to carefully design an information disclosure mechanism to incentivize each strategic labeler's truthful feedback of his preference. We should also ensure the final outcome of the system's evaluation to be as close to the actual $\sum_{i=1}^N w_i\theta_i$ as possible for efficiency of RLHF learning. We summarize these two desired properties for mechanism design:

- *Truthfulness.* A mechanism is truthful if each labeler i chooses the feedback $m_i(\theta_i)$ from his feedback space $\bar{M}_i$ that most accurately reveals his preference θ_i.
- *Efficiency.* A mechanism is efficient if it reduces the system's performance loss in (2) as much as possible.

3 Two Benchmark Schemes: RLHF Practice and Weighted Median

In this section, we analyze two common schemes used in the literature, which will serve as two benchmarks for our PAID mechanism to compare later.

3.1 Benchmark 1: Current Practice of RLHF

The current practice of RLHF (e.g., [23], [26],[28]) is to take an average of strategic labelers' preference feedback, ignoring their strategic manipulation for their own benefit. As a result, each strategic labeler reports an arbitrary preference instead of the ground-truth, making it meaningless for the system to use any. This is so called cheap-talk in [1] and [27]. Thus, the best the system can do is to blindly minimize its performance loss in (2) by determining the final evaluation $\hat{\theta}$ according to $\{\theta_i\}_{i=1}^N$ distributions as follows:

$$\min_{\hat{\theta}} \mathbf{E}_{\{\theta_i\}_{i=1}^N}\left(\sum_{i=1}^N w_i\theta_i - \hat{\theta}\right)^2. \quad (3)$$

As (3) is convex in $\hat{\theta}$, we can solve it by the first-order condition to determine a final evaluation $\hat{\theta}$. By substituting it into (2), we can obtain the system's performance loss.

Lemma 3.1. *At the equilibrium of benchmark 1, the LLM system determines the final evaluation $\hat{\theta}$ as a weighted sum of the mean μ_i*

[1] There is no strategic decision for human feedback or system evaluation in Stages 0 and III.

of each labeler i's preference to obtain performance loss as follows:

$$\hat{\theta} = \sum_{i=1}^{N} w_i \mu_i,\ L_1 = \sum_{i=1}^{N} w_i^2 \sigma_i^2. \tag{4}$$

Besides, if each labeler i has the same variance for preference θ_i and the same weight $w_i = \frac{1}{N}$ for any $i \in [N]$ of equal reputation, L_1 in (4) decreases with the labeler number N and tends to 0 only if $N \to \infty$.

At benchmark 1, the system can never learn each labeler i's preference θ_i. It is thus best to use the mean μ_i of each θ_i in the final preference decision to minimize its performance loss, which incurs a variance σ_i^2 from each $i \in [N]$. Besides, if each labeler i has the same θ_i variance and weight, as N increases, more data inputs from other labelers balance the difference between the weighted preference $\sum_{i=1}^{N} w_i\theta_i$ and the weighted mean in (4). The system thus incurs a smaller performance loss. Note that benchmark 1 needs a large number of labelers for a small performance loss, which is difficult to satisfy for RLHF. We are thus motivated to find some other schemes to reduce the system's performance loss.

3.2 Benchmark 2: Weighted Median Scheme

In the algorithmic game theory literature, the popular non-monetary "weighted median" scheme is widely used to motivate strategic labelers' truthful reporting, where the system commits to the weighted median of labelers' reports (e.g., [35]). We give the formal definition.

Definition 3.2 (Weighted Median Scheme). The system first reorganizes labelers' preference feedback $\{m_i(\theta_i)\}_{i=1}^{N}$ in an increasing order as $m_{j_1} \le \cdots \le m_{j_N}$. It then chooses the weighted median $\hat{\theta} = m_{j_n}$ as its final decision, where

$$n = \min\left\{k \middle| \sum_{i<k} w_{j_i} \le \frac{1}{2}, \sum_{i>k} w_{j_i} \le \frac{1}{2}, 1 \le k \le N.\right\}.$$

Recall that each labeler i's loss in (1) is singly-peaked at his preference θ_i. If his θ_i is less than weighted median m_{j_n}, his misreporting of $m_i < m_{j_n}$ does not alter m_{j_n}. If he misreports m_i to be larger than m_{j_n}, we have the new weighted median m'_{j_n} greater than m_{j_n}. His loss in (1) only increases due to $\theta_i < m_{j_n} \le m'_{j_n}$. Similarly, he will not misreport if $\theta_i \ge m_{j_n}$. We have the following.

Lemma 3.3. *Each labeler $i \in [N]$ truthfully feedbacks his preference under the weighted median scheme, i.e., $m_i(\theta_i) = \theta_i$.*

Now we are ready to analyze the system's performance loss performance in benchmark 2. Substitute the weighted median feedback as the final preference to (2), we derive the closed-form performance loss under the case of uniformly distributed $\{\theta_i\}_{i=1}^{N}$ below. Though more involved, the analysis of non-uniform distributions is similar, and we examine the system's performance loss for more general θ_i distributions later in Section 5.2.

Lemma 3.4. *Given each labeler i's preference $\theta_i \sim U[A, B]$ for $i \in [N]$, the system's performance loss in (2) under the weighted median scheme is*

$$L_2 = \sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{m^+\in\Phi_n} \frac{\mathbf{1}_{\sum_{i\in m^-} w_{j_i} \le \frac{1}{2}, \sum_{i\in m^+} w_{j_i} \le \frac{1}{2}}}{N \cdot C_{N-1}^{n-1}} \Bigg(\sum_{i\ne m} w_i^2\sigma^2 + \frac{3\sigma^2}{5}\Bigg(\sum_{i\ne m, i\in m^+} w_i \sum_{j\ne m, i; j\in m^+} w_j + \sum_{i\ne m, i\in m^-} w_i \sum_{j\ne m, i; j\in m^-} w_j \Bigg) + \frac{\sigma^2}{10}\Bigg(\sum_{i\ne m, i\in m^+} w_i \sum_{j\ne m, i; j\in m^-} w_j + \sum_{i\ne m, i\in m^-} w_i \sum_{j\ne m, i; j\in m^+} w_j \Bigg)\Bigg), \tag{5}$$

where $\mathbf{1}_{(\cdot)}$ is the indicator function, set $m^+ = \{i|\theta_i > \theta_m\}$, $m^- = \{i|\theta_i \le \theta_m, i \ne m\}$, $\Phi_n = \{S|S = \{i|\theta_i > \theta_m|\theta_m = \theta_{j_n}\}\}$ and j_n is the index of the n_{th} smallest feedback among $\{\theta_i\}_{i=1}^{N}$. If each labeler i has preference $\theta_i \sim U[A, B]$ and identical weight $w_i = \frac{1}{N}$ for any $i \in [N]$, the system's performance loss in (5) is simplified to[2]

$$L_2 = \begin{cases} \frac{(N-1)(7N+1)}{20N^2}\sigma^2, & \text{if } N \text{ is odd,} \\ \frac{7N^2-6N+4}{20N^2}\sigma^2, & \text{if } N \text{ is even.} \end{cases}$$

L_2 decreases as N increases from 2 to 3 and then increases with $N \ge 3$.

At benchmark 2, if each labeler has the same weight and his θ_i has the same uniform distribution, the system incurs a smaller performance loss as the labeler number N increases from 2 to 3. The reason is that with $N = 2$, both labelers are equally critical to the system's performance loss due to having the same weight. The system's choice of either to be the final decision results in a non-small loss. However, as N keeps increasing from 3, there are more labelers whose preferences are different from the weighted median one. This causes the system to incur much loss with non-median labelers according to (1). The system's performance loss increases and cannot be zero even for $N \to \infty$, which is different from benchmark 1.

With Lemmas 3.1 and 3.4, we are now ready to compare the system's performance loss under benchmarks 1 and 2 below for uniform distribution. We compare for other θ_i distributions later in Section 5.2 to show similar results.

Corollary 3.5. *Given each labeler i's preference $\theta_i \sim U[A, B]$ and the same weight $w_i = \frac{1}{N}$ for $i \in [N]$, the system in benchmark 2 incurs a smaller performance loss than 1 only if $N \in \{2, 3\}$, but incurs a larger performance loss if $N \ge 4$.*

With a small labeler number $N \in \{2, 3\}$, benchmark 2 incurs a smaller system loss than benchmark 1 due to its labelers' truthful feedback of preferences. However, as N keeps increasing from 3 and there are more non-median labelers, it incurs even a larger performance loss than benchmark 1. We are well motivated to develop a brand new mechanism to substantially reduce the system's performance loss for any N and preference distribution.

4 Our new mechanism Design: PAID

In this section, we design a new PArtial Information Disclosure (PAID) mechanism to substantially decrease the system's performance loss. We will compare its performance against the two benchmarks later in Section 5.

[2] We choose the $\lfloor \frac{N}{2} \rfloor$th smallest feedback as the median of N feedback.

4.1 Introduction of Our PAID Mechanism

According to our analysis of benchmark 2 in Section 3.2, it is too costly on the system side to ensure each labeler i's truthful feedback of precise θ_i by committing the weighted median scheme. Instead, we no longer expect each labeler to feedback precisely a point but a range of preference, by relaxing the feedback content for higher efficiency purpose. Note that each labeler i's loss $\ell_i(\theta_i)$ in (1) is singly-peaked at θ_i, implying that he may truthfully feedback his θ_i realization belonging to a range. As a new feedback method, we propose to carefully partition θ_i's distribution range $[A_i, B_i]$ into K_i sub-intervals and ask labeler i to tell which sub-interval his θ_i belongs to. We redefine such an labeler i's feedback space $\bar{M}_i$ instead of $[A_i, B_i]$ in the following.

Definition 4.1 (Labelers' New Feedback Space $\bar{M}_i$ in PAID). The system determines $\Theta_i=\{[\theta_{i,k-1}, \theta_{i,k}]\}_{k=1}^{K_i}$ as the set of K_i connected sub-intervals of labeler i's θ_i distribution interval $[A_i, B_i]$, where $\theta_{i,0} = A_i$, $\theta_{i,K_i} = B_i$, and $\{\theta_{i,k}\}_{k=0}^{K_i}$ is a strictly increasing sequence in k. It then invites each labeler i to feedback in space $\bar{M}_i = \{1, \cdots, K_i\}$, where labeler i's feedback $m_i = k$ indicates $\theta_i \in [\theta_{i,k-1}, \theta_{i,k}]$, $k \in \{1, \cdots, K_i\}$.

As each labeler i may have distinct $[A_i, B_i]$, we may partition differently. Further, the partitioned sub-intervals should be properly designed to ensure truthful feedback from each labeler in $\bar{M}_i$. Note that the RLHF practice of benchmark 1 is a special case of our PAID mechanism by setting $K_i = 1$ for any $i \in [N]$. Given Definition 4.1, we now introduce our PAID mechanism.

Definition 4.2 (PArtial Information Disclosure Mechanism PAID). The system designs the PAID mechanism as follows.

- Step 1: The system first determines the sub-interval number K_i and feedback sub-interval set $\Theta_i = \{[\theta_{i,k-1}, \theta_{i,k}]\}_{k=1}^{K_i}$ to each labeler $i \in [N]$, where $\{\theta_{i,k}\}_{k=1}^{K_i-1}$ for any $K_i \geq 2$ are solutions to

$$\Big(\theta_{i,k} - w_i \mathbb{E}[\theta_i | \theta_i \in [\theta_{i,k-1}, \theta_{i,k}]] - \sum_{j\neq i}^{N} w_j\mu_j\Big)^2 =$$

$$\Big(\theta_{i,k} - w_i \mathbb{E}[\theta_i | \theta_i \in [\theta_{i,k}, \theta_{i,k+1}]] - \sum_{j\neq i}^{N} w_j\mu_j\Big)^2, \tag{6}$$

  $k \in [K_i - 1]$, and $\mathbb{E}[\theta_i | \theta_i \in [\theta_{i,k-1}, \theta_{i,k}]]$ denotes the conditional mean of θ_i given that it belongs to the sub-interval of $[\theta_{i,k-1}, \theta_{i,k}]$.
- Step 2: Each labeler i chooses his feedback $m_i(\theta_i) = k$, $k \in \bar{M}_i$ to minimize his loss in (1), telling $\theta_i \in [\theta_{i,k-1}, \theta_{i,k}]$.
- Step 3: After receiving labelers' feedback $\{m_i(\theta_i)\}_{i=1}^N$, the system determines the final preference for $k \in [K_i]$:

$$\hat{\theta}^*(\{\hat{\theta}_i(m_i(\theta_i))\}_{i=1}^N) = \sum_{i=1}^{N} w_i \hat{\theta}_i(m_i(\theta_i)), \text{ where}$$

$$\hat{\theta}_i(m_i(\theta_i) = k) = \mathbb{E}[\theta_i | \theta_i \in [\theta_{i,k-1}, \theta_{i,k}]]. \tag{7}$$

Note that (6) and K_i are properly designed. The next proposition shows that our PAID mechanism in Definition 4.2 ensures each labeler i's truthful feedback of the sub-interval containing his θ_i, proved by the convexity of (1) in each θ_i.

Proposition 4.3. *Our PAID mechanism in Definition 4.2 is truthful such that each labeler i truthfully feedbacks:*

$$m_i^*(\theta_i) = \{k | \theta_i \in [\theta_{i,k-1}, \theta_{i,k}],\ 1 \leq k \leq K_i.\}.$$

4.2 Optimization of Feedback Sub-Internals $\{\theta_{i,k}\}_{k=1}^{K_i-1}$ for Our PAID Mechanism

To obtain clear engineering insights on our PAID mechanism, in the following, we first focus on the case of each labeler i's uniformly distributed preference, i.e., $\theta_i \in U[A_i, B_i]$. We relax this assumption to analyze for any continuous θ_i distribution later in Section 5.2.

We use backward induction to solve our PAID mechanism. In Step 3, upon receiving labelers' truthful feedback $\{m_i(\theta_i)\}_{i=1}^N$ from Step 2, the system determines each $\hat{\theta}_i(m_i(\theta_i) = k) = (\theta_{i,k-1} + \theta_{i,k})/2$ in (7). Substituting $\hat{\theta}_i(m_i(\theta_i))$ into (6), we derive that labeler i's sub-interval boundaries $\{\theta_{i,k}\}_{k=1}^{K_i-1}$ in Step 1 are the solutions to

$$\theta_{i,k+1} - \frac{2(2-w_i)}{w_i}\theta_{i,k} + \theta_{i,k-1} = -\frac{4\sum_{j\neq i} w_j\mu_j}{w_i}, \tag{8}$$

where $k \in [K_i - 1]$. As (8) is a linear, autonomous, second-order difference equation, we divide it into a steady-state part with $\theta_{i,k} = \bar{\theta}_i$ for all $k \geq 0$ and a homogeneous part:

$$\theta_{i,k+1} - \frac{2(2-w_i)}{w_i}\theta_{i,k} + \theta_{i,k-1} = 0,\ k \in [K_i - 1].$$

As the solution to (8) is the sum of the solutions to the steady-state and the homogeneous parts, we have the following.

Proposition 4.4. *Given each labeler i's sub-interval number K_i fixed in Step 1 of our PAID mechanism, the system determines his sub-interval boundaries in closed-form $\{\theta_{i,k}\}_{k=0}^{K_i}$ below:*

$$\theta_{i,k} = \alpha_i(K_i)\left(\frac{(1+\sqrt{1-w_i})^2}{w_i}\right)^k + \frac{\sum_{j\neq i} w_j\mu_j}{1-w_i} + \beta_i(K_i)\left(\frac{(1-\sqrt{1-w_i})^2}{w_i}\right)^k,\ 0 \leq k \leq K_i,\ i \in [N], \tag{9}$$

where

$$\alpha_i(K_i) = \frac{-\left(A_i - \frac{\sum_{j\neq i} w_j\mu_j}{1-w_i}\right)\left(\frac{(1-\sqrt{1-w_i})^2}{w_i}\right)^{K_i} + B_i - \frac{\sum_{j\neq i} w_j\mu_j}{1-w_i}}{\left(\frac{(1+\sqrt{1-w_i})^2}{w_i}\right)^{K_i} - \left(\frac{(1-\sqrt{1-w_i})^2}{w_i}\right)^{K_i}},$$

$$\beta_i(K_i) = \frac{\left(A_i - \frac{\sum_{j\neq i} w_j\mu_j}{1-w_i}\right)\left(\frac{(1+\sqrt{1-w_i})^2}{w_i}\right)^{K_i} + \frac{\sum_{j\neq i} w_j\mu_j}{1-w_i} - B_i}{\left(\frac{(1+\sqrt{1-w_i})^2}{w_i}\right)^{K_i} - \left(\frac{(1-\sqrt{1-w_i})^2}{w_i}\right)^{K_i}}.$$

Note that we have not determined the sub-interval number K_i for each labeler $i \in [N]$ yet. The system's objective is to choose the optimal $\{K_i^*\}_{i=1}^N$ to minimize its performance loss in (2), subject to labelers' truthful feedback. According to (2) and Proposition 4.4, we have the following result.

Corollary 4.5. *The system's performance loss in* (2) *of our PAID decreases with each sub-interval number K_i, $i \in [N]$.*

With a larger sub-interval number K_i, the system divides more sub-intervals for labeler i to feedback and can extract more accurate

information on his θ_i. For the special case of each labeler i's sub-interval number $K_i = 1$ for $i \in [N]$, the system's performance loss degenerates to L_1 in (4) of benchmark 1.

According to Corollary 4.5, the system's performance loss of our PAID decreases with each labeler i's sub-interval number K_i. We thus design each K_i to be as large as possible to minimize the system's performance loss. Note that the solutions in (9) cannot guarantee that $\{\theta_{i,k}\}_{k=0}^{K_i}$ is strictly increasing in k, which needs to be satisfied for ensuring each labeler i's truthful feedback. We have the following.

Proposition 4.6. *In our PAID mechanism, the optimal sub-interval number K_i^* for each labeler i's feedback space $\bar{M}_i$ to minimize its performance loss is given in closed form by:*

$$K_i^* = \begin{cases} \left\lfloor \log_{\frac{(1+\sqrt{1-w_i})^2}{w_i}} \frac{C_s + \sqrt{C_s^2 - 4\frac{(1+\sqrt{1-w_i})^2}{w_i}}}{2} \right\rfloor, & \text{if } \bar{\mu}_{-i} \le A_i, \\ \infty, & \text{if } \bar{\mu}_{-i} \in (A_i, B_i), \\ \left\lfloor \log_{\frac{(1+\sqrt{1-w_i})^2}{w_i}} \frac{C_b + \sqrt{C_b^2 - 4\frac{(1+\sqrt{1-w_i})^2}{w_i}}}{2} \right\rfloor, & \text{if } \bar{\mu}_{-i} \ge B_i, \end{cases} \tag{10}$$

where we have

$$\bar{\mu}_{-i} = \frac{\sum_{j\neq i}^{N} w_j \mu_j}{1 - w_i}, C_s = \left(\frac{(1+\sqrt{1-w_i})^2}{w_i} + 1\right)\frac{B_i - \bar{\mu}_{-i}}{A_i - \bar{\mu}_{-i}},$$
$$C_b = \left(\frac{(1+\sqrt{1-w_i})^2}{w_i} + 1\right)\frac{A_i - \bar{\mu}_{-i}}{B_i - \bar{\mu}_{-i}}.$$

For labeler i to truthfully feedback, he considers the effect of the other labelers' expected mean preference μ_{-i} in the aggregation in (10). If μ_{-i} is within his preference region, they are similar and it is easier for the system to persuade the labeler to follow large enough $K_i^* = \infty$ than the other cases of (10) to reveal more local information. However, if μ_{-i} is outside his region, they are very different and labeler i has the intention to manipulate his feedback to obviously affect the aggregation. As such, the system can only decide a finite K_i^* in (10) to ensure his truthful feedback in the rough sub-intervals.

Corollary 4.7. *Each labeler i's optimal sub-interval number K_i^* in (10) is non-increasing in his θ_i's lower bound A_i and is non-decreasing in the upper bound B_i, respectively. Besides, it increases with the others' expected mean $\bar{\mu}_{-i}$ in the range of $\bar{\mu}_{-i} < A_i$ and it decreases with $\bar{\mu}_{-i}$ in the range of $\bar{\mu}_{-i} > B_i$.*

With decreased A_i or increased B_i for a larger range $[A_i, B_i]$ for possible θ_i, the system tends to decide a larger sub-interval number K_i^* to narrow each sub-interval for better inference. As the others' expected mean $\bar{\mu}_{-i}$ increases to approach A_i, its disturbance on labeler i's truthful feedback is mitigated. The system can divide more sub-intervals for labeler i and K_i^* thus increases. However, as $\bar{\mu}_{-i}$ keeps increasing beyond B_i, its disturbance on labeler i's truthful feedback becomes intensified. The system can only divide fewer sub-intervals for labeler i and K_i^* thus decreases.

One may be curious about our PAID mechanism under $K_i^* \to \infty$ for $i \in [N]$ and wonders whether it reveals the actual θ_i information. According to Propositions 4.4 and 4.6, the following tells that it is not the case. It is still necessary to ensure each labeler's truthful feedback in the sub-intervals.

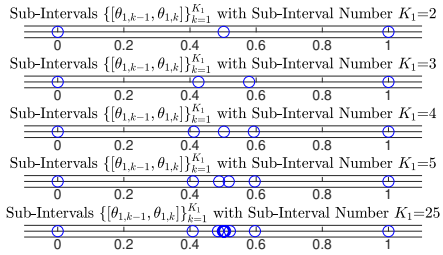

Figure 2: labeler 1's feedback sub-intervals $\{[\theta_{1,k-1},\theta_{1,k}]\}_{k=1}^{K_1}$ in (6) versus his sub-interval number K_1. Here we consider 4 labelers in total and each labeler i's θ_i follows a truncated normal distribution.

Proposition 4.8. *Given $\bar{\mu}_{-i} = \frac{\sum_{j\neq i}^{N} w_j\mu_j}{1-w_i}$ falls within (A_i, B_i) for any $i \in [N]$, the system optimally designs the sub-interval boundaries $\{\theta_{i,k}\}_{i=0}^{K_i^*\to\infty}$ for PAID as $\theta_{i,0}$=A_i, $\theta_{i,\infty}$=B_i and*

$$\theta_{i,k} = -(\bar{\mu}_{-i} - A_i)\left(\frac{(1-\sqrt{1-w_i})^2}{w_i}\right)^k + \bar{\mu}_{-i},\ k < \infty, \tag{11}$$

which ensures the first and the last sub-intervals do not vanish with $K_i^* \to \infty$:

$$\theta_{i,1} - \theta_{i,0} = (\bar{\mu}_{-i} - A_i)\left(1 - \frac{(1-\sqrt{1-w_i})^2}{w_i}\right) > 0,$$
$$\theta_{i,K_i^*} - \theta_{i,K_i^*-1} = B_i - \bar{\mu}_{-i} > 0.$$

Figure 2 considers an example of truthful aggregation from 4 strategic labelers, with each θ_i following a truncated normal distribution. Consider a particular labeler 1, as his K_1 increases, the system divides more sub-intervals around the mean 0.47 in a preference range of $[0, 1]$. Yet, the first and the last sub-intervals $[\theta_{1,0}, \theta_{1,1}]$ and $[\theta_{1,K_1-1}, \theta_{1,K_1}]$ do not vanish as K_1 increases to 25 and beyond, which is consistent with Proposition 4.8. This ambiguity helps hide labeler i's actual θ_i and ensures his truthful feedback of θ_i's sub-interval.

5 PAID Performance Analysis and Comparison with two benchmarks

Besides truthful property, in this section, we continue to show our PAID's efficiency advantages over the benchmark schemes in Section 3 to reduce the system's performance loss. We first make analytical comparisons in Section 5.1. In Section 5.2, we propose an efficient algorithm for our PAID design to fit more general θ_i distributions and provide numerical comparison with the two benchmarks.

5.1 Analytical Comparisons with The Two Benchmarks

Combining Corollary 4.5 and Proposition 4.6, we obtain the system's performance loss in (2) of our PAID mechanism below.

Lemma 5.1. *Given each labeler i's preference $\theta_i \in U[A_i, B_i]$ with the same mean $\mu_i = \mu$ for any $i \in [N]$, the system's performance loss*

of our PAID mechanism is

$$L^* = \sum_{i=1}^{N} \frac{(1-w_i)w_i^2}{4-w_i}\sigma_i^2. \tag{12}$$

Specifically, if each labeler $i \in [N]$ has the preference $\theta_i \sim U[A,B]$ in the same interval and has the same weight $w_i = \frac{1}{N}$, L^* in (12) decreases with $N \geq 2$ and tends to 0 as $N \to \infty$.

In the following, we first compare the system's performance loss of our PAID mechanism with benchmark 1.

Proposition 5.2. *Given each labeler i has the preference $\theta_i \in U[A_i,B_i]$ with the same mean $\mu_i = \mu$ for any $i \in [N]$, the ratio r_1 between the system's performance loss L^* in (12) of our PAID mechanism and L_1 in (4) of benchmark 1 is*

$$r_1 = \frac{L^*}{L_1} = \frac{\sum_{i=1}^{N} \frac{(1-w_i)w_i^2}{4-w_i}\sigma_i^2}{\sum_{j=1}^{N} w_j^2\sigma_j^2}, \tag{13}$$

which is always less than $\frac{1}{4}$. Besides, if each labeler $i \in [N]$ has the preference $\theta_i \sim U[A,B]$ and has the same weight $w_i = \frac{1}{N}$, we have r_1 in (13) increases with N and tends to $\frac{1}{4}$ as $N \to \infty$.

Our PAID mechanism substantially reduces the system's performance loss compared to benchmark 1 for any N. As benchmark 1 performs its best as $N \to \infty$, the ratio r_1 reaches $\frac{1}{4}$ then.

Now we compare the system's performance loss of our PAID with benchmark 2. We have the following.

Proposition 5.3. *If each labeler $i \in [N]$ has the preference $\theta_i \sim U[A,B]$ and the same weight $w_i = \frac{1}{N}$, the ratio r_2 between the system's performance loss L^* in (12) of our PAID and L_2 in (5) of benchmark 2 is*

$$r_2 = \frac{L^*}{L_2} = \begin{cases} \frac{20N}{(4N-1)(7N+1)}, & \text{if } N \text{ is odd,} \\ \frac{20(N-1)N}{(4N-1)(7N^2-6N+4)}, & \text{if } N \text{ is even,} \end{cases} \tag{14}$$

which is no more than $\frac{2}{7}$, decreases with N and tends to 0 as $N \to \infty$.

As the labeler number N increases, there are more labelers whose preference realizations are different from the weighted median and benchmark 2 worsens as shown in Lemma 3.4. This causes the ratio r_2 to decrease with N and our PAID's advantage over benchmark 2 becomes more obvious.

5.2 Performance Comparison for General θ_i Distributions

Recall that in Sections 4.2 and 5.1, we assume that each θ_i follows a uniform distribution. In this subsection, we relax this assumption to fit our PAID for generally distributed θ_i. We give a sufficient condition for ensuring labeler i's non-trivial sub-interval number K_i design below.

Lemma 5.4. *Given any continuous PDF $f_i(\cdot)$ of labeler i's preference $\theta_i \in [A_i,B_i]$, his feedback sub-interval number K_i in our PAID is no smaller than 2 as long as the following condition is satisfied:*

$$\frac{(2-w_i)A_i - w_i\mu_i}{2} < \sum_{j=1,j\neq i}^{N} w_j\mu_j < \frac{(2-w_i)B_i - w_i\mu_i}{2}.$$

Algorithm 1 Solving sub-interval boundaries $\{\{\theta_{i,k}\}_{k=0}^{K_i}\}_{i=1}^{N}$ in (6) for general $\{\theta_i\}_{i=1}^{N}$ distributions.

Input: Labeler number N, labelers' preference ranges $\{[A_i,B_i]\}_{i=1}^{N}$, PDFs $\{f_i(\cdot)\}_{i=1}^{N}$, labelers' maximum searching rounds $\{T_i\}_{i=1}^{N}$, labelers' preference means $\{\mu_i\}_{i=1}^{N}$, objective functions $\{\mathbf{h}_i\}_{i=1}^{N}$, and digit precisions $\{\epsilon_i\}_{i=1}^{N}$.

Output: Sub-interval boundaries $\{\{\theta_{i,k}\}_{k=0}^{K_i^{max}}\}_{i=1}^{N}$.

1: **Initialization**: $K_i \leftarrow 2$, $\mathbf{s}_i^0 \leftarrow (s_{i,1}^0, \cdots, s_{i,K_i-1}^0)$, $t \leftarrow 0$, $\mathbf{\Gamma}_1 \leftarrow \emptyset$, $\theta_{i,0} \leftarrow A_i$, $\theta_{i,K_i} \leftarrow B_i$.
2: Update $\mathbf{s}_i^{t+1} \leftarrow \mathbf{s}_i^t - \mathbf{J}_i(\mathbf{s}_i^t)^{-1} * \mathbf{h}_i(\mathbf{s}_i^t)$ until $t \geq T_i$ or $\|\mathbf{s}_i^t - \mathbf{s}_i^{t-1}\|_2 \leq 10^{-\epsilon_i}$. Use the pseudo-inverse of $\mathbf{J}_i(\mathbf{s}_i^t)$ if $\mathbf{J}_i(\mathbf{s}_i^t)^{-1}$ does not exist.
3: **if** $\exists t < T_i$ such that $\|\mathbf{s}_i^{t+1} - \mathbf{s}_i^t\|_2 \leq 10^{-\epsilon_i}$ **then**
4: $\quad \mathbf{\Gamma}_{K_i} \leftarrow \mathbf{s}_i^{t+1}$ and repeat Steps 1-3 with $K_i \leftarrow K_i + 1$.
5: **else**
6: $\quad K_i^{max} \leftarrow K_i - 1$, $\{\theta_{i,k}\}_{k=0}^{K_i^{max}} \leftarrow (A_i, \mathbf{\Gamma}_{K_i^{max}}, B_i)$.
7: Repeat Steps 1-6 for all labelers $i \in [N]$.

Lemma 5.4 is similar to Proposition 4.6 for the case of uniformly distributed θ_i. Recall that K_i=1 reveals no information and degenerates to benchmark 1. If the others' weighted mean $\sum_{j=1,j\neq i}^{N} w_j\mu_j$ lies in a range close to labeler i's θ_i realization range of $[A_i,B_i]$ as above, each labeler i worries less about the aggregation disturbance from the other labelers, and is easier to be persuaded by the system to feedback truthfully in more partitioned sub-intervals to reveal preference information.

In the following, we introduce an efficient algorithm in Algorithm 1 to solve each labeler i's truthful sub-interval boundaries $\{\theta_{i,k}\}_{k=0}^{K_i}$ for any θ_i distribution, by solving a set of equations in (6) and determining the maximum possible K_i called K_i^{max}. Let us first introduce some essential notations. Define vectors $\mathbf{s}_i^j = (s_{i,1}^j, \cdots, s_{i,K_i-1}^j)$ and objective functions $\mathbf{h}_i := (h_{i,1}, \cdots, h_{i,K_i-1})$ such that (6) can be rewritten as $h_{i,k}(\theta_{i,k}) = 0$. Denote $\mathbf{J}_i(\mathbf{s})$ as the Jacobian matrix of $\mathbf{h}_i(\mathbf{s})$. Our objective is to numerically find a solution $\mathbf{s}$ of $\mathbf{h}_i$ such that $\mathbf{h}_i(\mathbf{s}) = 0$. Inspired by Corollary 4.5, for better system performance, we design Algorithm 1 to initiate with a sub-interval number of $K_i = 2$ and stop when monotonicity of $\theta_{i,k}$ sequence in (6) almost fails, obtaining the largest feasible K_i^{max} that supports truthful feedback for labeler i. The computational complexity of Algorithm 1 is

$$O\Big(\sum_{i=1}^{N}\sum_{K_i=2}^{K_i^{max}} K_i^3 \cdot \log_2 \epsilon_i\Big),$$

where N denotes the labeler number, K_i^{max} denotes labeler i's maximum possible sub-interval number, and ϵ_i is the digit precision of labeler i. Thus, Algorithm 1 incurs a polynomial complexity order in each labeler i's maximum possible sub-interval number K_i^{max}. It is also scalable with large labeler number N. Additionally, Algorithm 1 exhibits good convergence with a quadratic rate.

Proposition 5.5. *Suppose that labeler i's preference θ_i follows a truncated normal distribution. Denote $\boldsymbol{\theta}_i^*$ as a root of $\mathbf{h}_i(\boldsymbol{\theta}_i^*) = 0$ in*

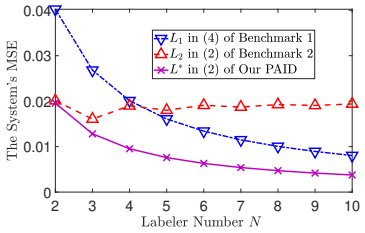

Figure 3: The system's performance loss L_1 in (4) of benchmark 1, L_2 in (2) using benchmark 2 and L^* in (2) using our PAID versus the labeler number N. Here we consider that each labeler i's preference θ_i follows a truncated normal distribution and he has the same weight of $w_i = \frac{1}{N}$.

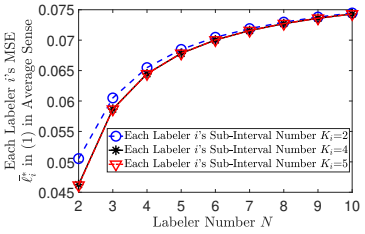

Figure 4: Each labeler i's performance loss $\bar{\ell}_i$ in (1) in average sense of our PAID versus the labeler number N and each labeler i's sub-interval number K_i, respectively. Here we consider that each labeler i's preference θ_i follows a truncated normal distribution and he has the same weight of $w_i = \frac{1}{N}$.

Algorithm 1. There exists a positive $\delta > 0$ such that if $||\mathbf{s}_i^0 - \boldsymbol{\theta}_i^*|| < \delta$ holds, we have

$$\lim_{t\to\infty} \frac{||\mathbf{s}_i^{t+1} - \boldsymbol{\theta}_i^*||}{||\mathbf{s}_i^t - \boldsymbol{\theta}_i^*||} = 0.$$

Besides, there exists a positive $M > 0$ such that for all $t \geq 0$,

$$||\mathbf{s}_i^{t+1} - \boldsymbol{\theta}_i^*|| \leq M||\mathbf{s}_i^t - \boldsymbol{\theta}_i^*||^2.$$

In the following, we numerically analyze and compare the performance of our PAID and the two benchmarks for truncated normal distributions of labelers' preferences.

Figure 3 shows the system's performance loss L_1 in (4) of benchmark 1, L_2 in (2) using benchmark 2 and L^* in (2) using our PAID versus the labeler number N, respectively. Figure 3 implies that the system always incurs a much smaller performance loss under our PAID than the two benchmarks, which is consistent with Propositions 5.2 and 5.3. Besides, the system first incurs a larger performance loss under benchmark 1 than 2 if the labeler number is small as $N \leq 4$, and then incurs a smaller performance loss if $N \geq 5$, consistent with Corollary 3.5.

Figure 4 shows each labeler i's performance loss in (1) in average sense of our PAID versus the labeler number N and labelers' sub-interval number K, respectively. As N increases, each labeler expects more disturbance on aggregation by the others' preference realizations and his performance loss in average sense then increases. Perhaps surprisingly, like the system loss in Corollary 4.5, each labeler i's performance loss in average sense also decreases with his sub-interval number K_i. The system can learn each labeler i's θ_i more precisely, and in return the local model of the labeler is better reflected in the final inference. Figure 4 with minor curve difference for larger K_i also implies that the system does not need to implement many sub-intervals due to the diminishing performance gain.

6 Extension to Consider inaccurate human belief at the system

Recall that in Section 2, we assume that the system knows the probability distribution $f_i(\cdot)$ of each labeler i's preference $\theta_i \in [A_i, B_i]$, $i \in [N]$. In this section, we extend our PAID to address the more challenging scenario where the system holds inaccurate human belief of each labeler i's preference distribution. After imposing an inaccurate belief to the system, now θ_i's distribution changes from a deterministic range $[A_i, B_i]$ to $[A_i - n_i, B_i + n_i]$, where $n_i \in [-D, D]$ is the random noise and is unknown to the system. Without much loss of generality, we still suppose $A_i + B_i = 2\mu_i$ and bound the maximum deviation $D \in (\frac{A_i - B_i}{2}, \frac{B_i - A_i}{2})$. We focus on the case that PDF $f_i(\cdot)$ of each θ_i is symmetric, defined below.

Definition 6.1. A PDF $f_i(\cdot)$ is defined as symmetric if it satisfies $f(\mu_i - x) = f(\mu_i + x)$ around mean μ_i for any $x \in \mathcal{R}$.

Such inaccurate belief n_i on $[A_i - n_i, B_i + n_i]$ introduces difficulty to our prior PAID's feedback partition approach into sub-intervals. Our PAID design is still to incentivize labelers' truthful feedback of the sub-intervals containing their preferences. Our alternative idea is to determine each labeler i's sub-interval boundaries $\{\theta_{i,k}\}_{k=0}^{K_i}$ of our PAID according to his θ_i's largest possible range of $[A_i - D, B_i + D]$. With a similar analysis in Section 4, we have the following.

Proposition 6.2. *Suppose that each labeler i's preference θ_i follows a symmetric PDF $f_i(\cdot)$ in a range of $[A_i - n_i, B_i + n_i]$ with mean μ_i and unknown $n_i \in [-D, D]$ for $i \in [N]$. The system optimally determines each labeler i's sub-interval boundaries to be $\theta_{i,0} = A_i - D$, $\theta_{i,K_i} = B_i + D$ and $\{\theta_{i,k}\}_{k=1}^{K_i^*-1}$ as the solutions to*

$$(\theta_{i,k} - w_i E[\theta_i | \theta_i \in [\theta_{i,k-1}, \theta_{i,k}]] - \sum_{j\neq i} w_j \mu_j)^2$$
$$= (\theta_{i,k} - w_i E[\theta_i | \theta_i \in [\theta_{i,k}, \theta_{i,k+1}]] - \sum_{j\neq i} w_j \mu_j)^2,$$

where $k \in [K_i^* - 1]$ and K_i^* is optimally decided as labeler i's largest sub-interval number such that $\{\theta_{i,k}\}_{i=0}^{K_i^*}$ is a strictly increasing sequence. Each labeler i will truthfully feedback the index of the sub-interval containing his θ_i:

$$m_i^*(\theta_i) = \{k | \theta_i \in [\theta_{i,k-1}, \theta_{i,k}],\ k \in [K_i^*].\}.$$

To tell the effect of the system's inaccurate human belief, we define the system's performance loss increment ΔL^* caused by uncertain bounds of labeler i's θ_i as follows:

$$\Delta L^* = \mathbb{E}_{\{\theta_i\}_{i=1}^N, \{n_i\}_{i=1}^N}[L^*] - \mathbb{E}_{\{\theta_i\}_{i=1}^N}[L^* | n_i = 0]. \quad (15)$$

With a similar analysis in Section 4, the following proposition indicates that ΔL^* is limited in a quadratic term of D.

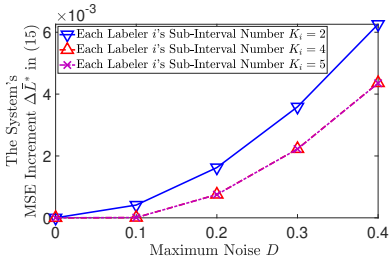

Figure 5: The system's performance loss increment ΔL^* in (15) versus the maximum noise D and labeler i's sub-interval number K_i, respectively. Here we consider N=6 and each θ_i follows a truncated normal distribution.

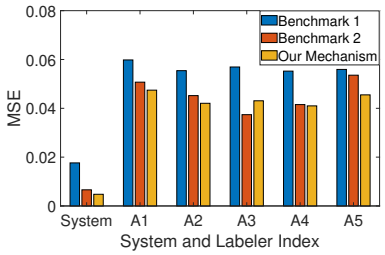

Figure 6: The system's MSE in (2) and each labeler's MSE in (1) under benchmarks 1, 2, and our PAID by using the TL;DR dataset, respectively.

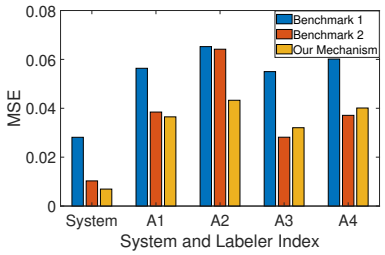

Figure 7: The system's MSE in (2) and each labeler's MSE in (1) under benchmarks 1, 2, and our PAID by using the CNN/DM dataset, respectively.

Proposition 6.3. *Given each labeler i's preference $\theta_i \sim U[A_i - n_i, B_i + n_i]$ and any realized noise $n_i \sim U[-D, D]$ for any $i \in [N]$, the system's performance loss increment ΔL^* in (15) is limited as $O(D^2)$.*

Figure 5 shows the system's performance loss increment ΔL^* in (15) versus the maximum noise D and labeler i's sub-interval number K_i, respectively. It indicates as the maximum noise D increases from 0 to 0.4, the system's maximum loss is just around 0.006, which is consistent with Proposition 6.3. Further, the system incurs a smaller performance loss increment as the sub-interval number increases from K_i=2 to 5, consistent with Corollary 4.5 and our K_i^* design in Proposition 6.2.

7 Case Study: Human-Annotated Text Summary in RLHF

In this section, we use two renowned datasets (including reddit post summaries and CNN article summaries) for human-annotated text summary in the RLHF literature to further verify our PAID's advantage over the benchmarks. Following the RLHF literature (e.g., [24], [30]), we evaluate the system's aggregation accuracy, which directly impacts the performance of the final fine-tuned policy.

7.1 Experiments Using The TL;DR Dataset for Post Summary

First, we follow [21] to use the typical TL;DR dataset in the RLHF literature (e.g., [23], [34]) for realistic performance evaluation, which contains around 123,169 posts from reddit.com across a variety of topics, as well as summaries of the posts written by the original poster. We further filter the TL;DR dataset to select 400 posts with understandable and clear summaries to ensure high quality of human feedback, where each summary is rated by 5 qualified labelers/groups in the range of 1 to 7. To make it consistent with our system model, we normalize each rate in the range of [0, 1]. We empirically approximate the PDF of each labeler's preference by his total 400 rates, which is then used to determine the system's final evaluation decision of human preferences under benchmark 1. For our PAID mechanism, each labeler i's sub-intervals $\{[\theta_{i,k-1}, \theta_{i,k}]\}_{k=1}^{K_i^*}$ in (6) with sub-interval number K_i^* of our PAID are computed and returned by Algorithm 1, which are then used to determine the system's final evaluation.

Figure 6 compares the system's MSE L in (2) and each labeler i's MSE ℓ_i in (1) under our PAID against the two benchmarks by using the filtered TL;DR dataset. Recall that benchmark 1 denotes the current practices of RLHF (e.g., [23], [26],[28]) and benchmark 2 denotes popular weighted median against strategic labelers (e.g., [35]). It indicates that our PAID incurs over 70% MSE decrement than benchmark 1 and 30% than benchmark 2, respectively. Since each labeler's preference realization can be distant from the weighted mean or median preference, he incurs a larger MSE from the final evaluation than the system.

Figure 6 further shows that our PAID incurs a smaller MSE for most labelers (i.e., 1, 2, 4, 5) than the two benchmarks. Recall in benchmark 1, the system blindly averages weighted means of five labelers' feedback as the final preference. Thus, each labeler's MSE does not vary much from the others'. At median benchmark 2, the system here chooses labeler 3's preference as the final evaluation. This leads to its smallest MSE by sacrificing all the others', especially for labelers 1 and 5. Differently, our PAID learns each labeler's preference as precisely as possible and fairly averages them in the final evaluation. Each thus obtains a smaller MSE than benchmark 1 and most obtain smaller MSE than benchmark 2.

7.2 Experiments Using The CNN/DM Dataset for Article Summary

We further follow [21] to use the typical CNN/DM dataset in the RLHF literature (e.g., [34], [23]) for realistic performance evaluation, which contains 2,500 articles in CNN and Daily Mail websites and the corresponding summaries. We further filter the CNN/DM

dataset to select 700 articles with understandable and clear summaries to ensure high quality of human feedback, where each summary is rated by 4 qualified labelers/groups in the range of 1 to 7. We follow the same procedure as in Section 7.1 to normalize the rates in th range of [0, 1] and obtain the preference PDFs and each labeler i's sub-intervals for determining the system's final evaluation of the benchmarks and our PAID mechanism, respectively.

Figure 7 compares the system's MSE L in (2) and each labeler i's MSE ℓ_i in (1) under our PAID against the two benchmarks by using the filtered CNN/DM dataset. It indicates that our PAID incurs over 70% MSE decrement than benchmark 1 and 30% MSE decrement than benchmark 2, respectively. Further, Figure 7 has the same insights as Figure 6 to benefit many labelers (e.g., 1 and 2) than the two benchmarks.

8 Conclusion

In this paper, we study how to truthfully aggregate human feedback from strategic labelers in RLHF. We design a novel PAID mechanism to ensure labelers' truthful feedback and minimize the system's performance loss. We prove that PAID's performance loss is at most 1/4 of the RLHF practice and 2/7 of the popular weighted median scheme, respectively. We also show that our PAID is robust to inaccurate human belief held by the system to yield limited loss. Finally, we run experiments using several real-world datasets to demonstrate our PAID's great advantages over the common benchmarks and show its positive "side-effect" to even align with most strategic labelers' preferences.

References

[1] Nemanja Antić and Nicola Persico. 2023. Equilibrium selection through forward induction in cheap talk games. *Games and Economic Behavior* 138 (2023), 299–310.

[2] Itai Arieli, Ivan Geffner, and Moshe Tennenholtz. 2023. Mediated cheap talk design. In *Proc. AAAI*, Vol. 37. 5456–5463.

[3] Mahsa Asadi, Aurélien Bellet, Odalric-Ambrym Maillard, and Marc Tommasi. 2022. Collaborative Algorithms for Online Personalized Mean Estimation. *Transactions on Machine Learning Research Journal* (2022).

[4] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. 2024. MaxMin-RLHF: Alignment with Diverse Human Preferences. In *ICML*.

[5] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement biases. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (2024).

[6] Yiding Chen, Jerry Zhu, and Kirthevasan Kandasamy. 2024. Mechanism Design for Collaborative Normal Mean Estimation. *NeuIPs* 36 (2024).

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

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

[9] Gabriel V Cruz Jr, Yunshu Du, and Matthew E Taylor. 2017. Pre-training neural networks with human demonstrations for deep reinforcement learning. *arXiv preprint arXiv:1709.04083* (2017).

[10] Zibin Dong, Yifu Yuan, Jianye Hao, Fei Ni, Yao Mu, Yan Zheng, Yujing Hu, Tangjie Lv, Changjie Fan, and Zhipeng Hu. 2023. Aligndiff: Aligning diverse human preferences via behavior-customisable diffusion model. *arXiv preprint arXiv:2310.02054* (2023).

[11] Avinava Dubey, Zhe Feng, Rahul Kidambi, Aranyak Mehta, and Di Wang. 2024. Auctions with llm summaries. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 713–722.

[12] Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. 2024. Mechanism design for large language models. In *Proceedings of the ACM on Web Conference 2024*. 144–155.

[13] Sharad Goel, Daniel M Reeves, and David M Pennock. 2009. Collective revelation: A mechanism for self-verified, weighted, and truthful predictions. In *Proceedings of the 10th ACM conference on Electronic commerce*. 265–274.

[14] Shugang Hao and Lingjie Duan. 2024. To Save Mobile Crowdsourcing from Cheap-talk: A Game Theoretic Learning Approach. *IEEE Transactions on Mobile Computing* (2024).

[15] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*. 64–67.

[16] Bernhard Jaeger, Andreas Geiger, et al. 2024. An invitation to deep reinforcement learning. *Foundations and Trends® in Optimization* 7, 1 (2024), 1–80.

[17] Majid Karimi and Nima Zaerpour. 2022. Put your money where your forecast is: Supply chain collaborative forecasting with cost-function-based prediction markets. *European Journal of Operational Research* 300, 3 (2022), 1035–1049.

[18] Yuqing Kong and Grant Schoenebeck. 2018. Eliciting expertise without verification. In *ACM EC*. 195–212.

[19] Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems* 36 (2023), 47669–47681.

[20] Jiaqian Li, Minming Li, and Hau Chan. 2024. Strategyproof Mechanisms for Group-Fair Obnoxious Facility Location Problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 9832–9839.

[21] Fei Liu et al. 2020. Learning to summarize from human feedback. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[22] Chris Lu, Timon Willi, Alistair Letcher, and Jakob Nicolaus Foerster. 2023. Adversarial cheap talk. In *International Conference on Machine Learning*. PMLR, 22917–22941.

[23] Long Ouyang and et al. 2022. Training language models to follow instructions with human feedback. *NeuIPs* 35 (2022), 27730–27744.

[24] Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman E Ozdaglar. [n. d.]. RLHF from Heterogeneous Feedback via Personalization and Preference Aggregation. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*.

[25] Silviu Pitis and Michael R Zhang. 2020. Objective social choice: Using auxiliary information to improve voting outcomes. *arXiv preprint arXiv:2001.10092* (2020).

[26] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Proc. NeuIPs* 36 (2023).

[27] Aleksei Smirnov and Egor Starkov. 2024. Designing Social Learning. *arXiv preprint arXiv:2405.05744* (2024).

[28] Feifan Song and et al. 2024. Preference ranking optimization for human alignment. In *Proc. AAAI*, Vol. 38. 18990–18998.

[29] Ermis Soumalias, Michael Curry, and Sven Seuken. 2024. Truthful Aggregation of LLMs with an Application to Online Advertising. In *Agentic Markets Workshop at ICML 2024*. https://openreview.net/forum?id=Pp6483Ma1m

[30] Haoran Sun, Yurong Chen, Siwei Wang, Wei Chen, and Xiaotie Deng. 2024. Mechanism Design for LLM Fine-tuning with Multiple Reward Models. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

[31] H Touvron and et al. 2023. Open and efficient foundation language models. *Preprint at arXiv. https://doi. org/10.48550/arXiv* 2302 (2023).

[32] Xuezhen Tu, Kun Zhu, Nguyen Cong Luong, Dusit Niyato, Yang Zhang, and Juan Li. 2022. Incentive mechanisms for federated learning: From economic and game theoretic perspective. *IEEE transactions on cognitive communications and networking* 8, 3 (2022), 1566–1593.

[33] Hristos Tyralis and Georgia Papacharalampous. 2024. A review of predictive uncertainty estimation with machine learning. *Artificial Intelligence Review* 57, 4 (2024), 94.

[34] Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*. 59–63.

[35] Ying Wang, Houyu Zhou, and Minming Li. 2024. Positive Intra-Group Externalities in Facility Location. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 1883–1891.

[36] Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025* (2023).

[37] Yutong Wu, Ali Khodabakhsh, Bo Li, Evdokia Nikolova, and Emmanouil Pountourakis. 2024. Eliciting truthful reports with partial signals in repeated games. *Theoretical Computer Science* 988 (2024), 114371.

[38] Wanqi Xue, Bo An, Shuicheng Yan, and Zhongwen Xu. 2023. Reinforcement learning from diverse human preferences. *arXiv preprint arXiv:2301.11774* (2023).

[39] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2024).

[40] Tianchen Zhu, Yue Qiu, Haoyi Zhou, and Jianxin Li. 2024. Decoding Global Preferences: Temporal and Cooperative Dependency Modeling in Multi-Agent Preference-Based Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17202–17210.