RLHF-Enhanced LLM Fine-Tuning for Cooperative Spectrum Sensing

Shugang Hao, *Member, IEEE,* and Lingjie Duan, *Senior Member, IEEE*

***Abstract*—Traditional cooperative spectrum sensing (CSS) engages a group of secondary users (SUs) to collaboratively detect available spectrum resources owned by primary users (PUs). However, it requires a precise model about channel or SU sensing performance to decide. To adapt to unknown or changing environments, we are the first to apply reinforcement learning from human feedback (RLHF)—a widely used fine-tuning technique in large language models (LLMs)—to enhance CSS. We formulate a new dynamic Bayesian game to model iterative interaction between the fusion center and SUs in dynamic CSS. We prove that no matter the prior model-based approach (e.g., k-out-of-N) or the current RLHF using uniform weighting incurs a non-vanishing regret of $\mathcal{O}(T)$ over a time horizon T. As such, our new RLHF iteratively weighs and chooses SUs in the fusion center based on their prior sensing performances. This approach not only incentivizes truthful reporting from SUs but also achieves a sublinear regret $\mathcal{O}(\sqrt{T})$ over time. Furthermore, we extend our RLHF design to another practical setting with limited SU feedback per time slot. Simulation results validate significant performance gains of our proposed mechanisms compared to benchmark schemes.**

***Index Terms*—Cooperative spectrum sensing, LLM fine-tuning, reinforcement learning from human feedback, online weighted aggregation, truthful mechanism design, regret analysis.**

I. Introduction

Though enhanced by exploring unused spectrum at higher frequencies, network operators still face limited spectrum resources due to the overwhelming number of network devices (e.g., [2], [3]). To improve spectrum utilization, cooperative spectrum sensing (CSS) assigns unlicensed secondary users (SUs) to opportunistically explore the licensed spectrum holes, which are originally allocated to licensed primary users (PUs) (e.g., [4]). Current CSS implementations commonly rely on the k-out-of-N scheme to commit to the sensing feedback that k SUs favor (e.g., [5]–[7]). However, such traditional CSS schemes require a precise model about channel or SU sensing performance to decide. This makes them difficult to generalize to unknown or changing network environments.

Recently, the communication society has explored deploying large language models (LLMs) for network control decisions in dynamic environments (e.g., semantic communication [8] and power control [9], [10]). To better meet application-specific user demands, pre-trained LLMs are fine-tuned using task-oriented datasets (e.g., [11]). Reinforcement learning from human feedback (RLHF) offers a promising way to align LLMs with practical human needs (e.g., [12], [13]). This approach collects online human feedback to build a preference dataset, which then trains and updates the learning policy.

To our best knowledge, we are the first to leverage RLHF for dynamic CSS. Each SU first obtains results about PU presence based on his local PU signal measurements. He then reports his sensing results as the online human feedback to the fusion center, which constructs an SU sensing dataset to train and update the LLM's CSS policy for iteratively identifying and prioritizing the most reliable SUs over time. Unlike traditional model-based CSS methods, this RLHF-aided CSS does not require a precise model about channel or SU sensing performance to decide. Further, it eliminates the need for extensive human-labeled datasets during policy training and updates, enabling data-efficient fine-tuning and robust adaptation to dynamic environments.

However, applying RLHF to CSS faces critical challenges. First, each SU senses the spectrum from a distinct physical location under varying channel conditions, leading to divergent (and possibly biased) results about spectrum availability. This makes it challenging for the fusion center to identify the most accurate SU and improve sensing accuracy in dynamic CSS. Second, as the fusion center hires SUs for CSS, they are non-cooperative to have their own selfish objectives as receiving the maximum credits or rewards from the fusion center. This can prevent them from sharing truthful sensing feedback to the fusion center (e.g., [4], [14], [15]), further complicating the fusion center's ability to learn reliable SUs. Current RLHF practices simply average human feedback during aggregation (e.g., [16]–[18]). This leads to our first key question:

- *Q1. How does prior model-based approach (e.g., the k-out-of-N rule) perform in the unknown or changing environment? Can we directly apply the current RLHF with uniform weighting?*

Later we prove that both the the k-out-of-N and the average feedback schemes do not track the most accurate SU in dynamic CSS, incurring a non-vanishing regret of $\mathcal{O}(T)$ in sensing accuracy over T time slots. As there is also a lack of study to address selfish reporting issues in recent online RLHF literature (e.g., [12], [13]), we turn to check if conventional wisdoms in CSS are applicable to our problem.

To address the second challenge, we find that game-theoretic studies on selfish SUs (e.g., [14], [15], [19], [20]) typically assume the binary decision of SUs, failing to solve our RLHF-based CSS with SUs' continuous sensing feedback. Relevant studies in facility location games (e.g., [21]) address truthful reporting incentives, where customers may strategically misreport locations to influence facility placement. While the median scheme (e.g., [22], [23]) effectively ensures truthfulness in these settings, we prove that it is no longer truthful for our

Part of this work has appeared in IEEE ICASSP 2025 [1].

S. Hao and L. Duan are with the Pillar of Engineering Systems and Design, Singapore University of Technology and Design, Singapore, 487372 Singapore. E-mail: shugang_hao@sutd.edu.sg, lingjie_duan@sutd.edu.sg.

problem and also incurs a non-vanishing regret similar to the average feedback scheme. Our second question is thus:

- *Q2. How to design an efficient and truthful mechanism for dynamic RLHF-based CSS with selfish SUs?*

A natural approach is to implement weighted aggregation and dynamically adjust each SU's weight in RLHF-based CSS. However, motivating SUs' truthful sensing feedback for a vanishing regret remains challenging. SUs' hidden and time-varying sensing results make it difficult for the fusion center to verify and correct misreports. Since the most accurate SU is unknown in the dynamic CSS process, it is difficult to dynamically weigh each SU for a vanishing regret.

We summarize our key novelty and main results below.

- *RLHF-enhanced LLM fine-tuning for cooperative spectrum sensing:* To our best knowledge, we are the first to leverage RLHF for dynamic CSS. Each SU reports his sensing results as the online human feedback to the fusion center, which constructs an SU sensing dataset to train and update the LLM's CSS policy for iteratively identifying and prioritizing the most reliable SUs over time. Unlike traditional model-based CSS methods (e.g., k-out-of-N [5]–[7]), our approach does not require a precise model about channel or SU sensing performance to decide. Distinct from the RLHF literature (e.g., [12], [13]), we address strategic misreporting by diverse SUs in dynamic CSS. We aim for new analytical studies to guide *how a fusion center can identify the most accurate SU over time in an RLHF-based dynamic CSS.*
- *Non-vanishing regrets of current practices:* The current practice of CSS widely adopts a k-out-of-N scheme to commit to the sensing feedback that k SUs favor. We prove that it fails to identify the most accurate SU in our problem and incurs a non-vanishing regret $\mathcal{O}(T)$ in sensing accuracy over T time slots. Similarly, we prove that the current RLHF practice of averaging user feedback per time slot also yields a non-vanishing regret $\mathcal{O}(T)$.
- *Our novel truthful online weighted aggregation mechanism:* We first formulate a new dynamic Bayesian game to model strategic interaction between the fusion center and SUs in dynamic CSS. We then propose a novel truthful online weighted aggregation mechanism within the RLHF framework to dynamically adjust SUs' weights in the sensing feedback aggregation based on their feedback accuracy during online learning. We prove that our mechanism guarantees truthful sensing feedback from SUs and achieves a sublinear regret $\mathcal{O}(\sqrt{T})$ over T time slots.
- *Extension to limited SU feedback:* In practice, waiting for multiple SUs to sense channels and report their measurements can introduce significant delays. Moreover, simultaneous feedback transmissions from multiple SUs may lead to overhead or signal collisions (e.g., [24], [25]). We then extend our analysis to a challenging limited SU feedback case where only one SU's sensing feedback is available per time slot. We propose a novel online mixed selection mechanism to ensure truthful feedback from any strategic SU while maintaining a sublinear regret $\mathcal{O}(\sqrt{T})$. Simulation results validate the significant performance gains of our proposed mechanisms compared to benchmark schemes.

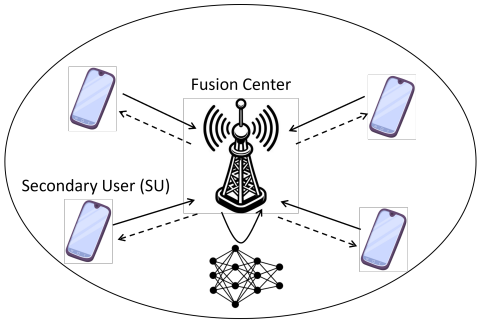

Fig. 1. System model of RLHF-based dynamic cooperative spectrum sensing. During each time slot $t \in [T]$, in Stage I, each SU first strategically reports his sensing results of PU presence at queried frequency bands. In Stage II, the fusion center aggregates SUs' sensing feedback for updating the RLHF policy. In Stage III, the fusion center adjusts each SU i's weight w_i^{t+1} according to his sensing feedback accuracy for the next time slot $t+1$'s iteration.

The rest of this paper is organized as follows. Section II introduces the system model and the dynamic Bayesian game formulation for RLHF-based dynamic CSS. Section III analyzes three common schemes used in the literature as benchmarks for our mechanism to compare later. Section IV details our proposed mechanism design and analysis. Section V extends the framework to limited SU feedback. Section VI presents simulation results. Section VII finally concludes.

II. System Model and Problem Formulation

In Section II-A, we introduce our system model. In Section II-B, we formulate a new dynamic Bayesian game and give desired properties for guiding our late mechanism design.

A. System Model of RLHF-Based CSS

As shown in Fig. 1, we focus on a secondary cognitive-radio network, containing a single fusion center and N SUs. The fusion center coordinates SUs' CSS and their access to a licensed PU spectrum. Our RLHF-based CSS starts from fine-tuning a pre-trained language model with supervised-learning approaches based on a CSS-oriented dataset to obtain a reference policy. Then, the fusion center iterates the RLHF process in T time slots (e.g., [12], [13]) for SUs' sensing feedback aggregation and RLHF policy update, where each time slot $t \in [T] := \{1, \cdots, T\}$ contains following three stages.

1) Stage I. Online Human Feedback: The fusion center selects m_t frequency bands for SUs to sense and constructs m_t prompts to query the LLM regarding PU presence at each band $j \in [m_t]$. Each prompt x_j^t contains historical PU activity, predefined PU occupancy time, detection threshold, and historical SU feedback on band j. For each prompt x_j^t, the LLM generates pairwise responses $(y_{l_j}^t, y_{l'_j}^t | x_j^t)$, indicating whether PUs are present (i.e., $y_{l_j}^t$) or absent (i.e., $y_{l'_j}^t$).

Concurrently, the fusion center collects SUs' sensing results of PU presence at the queried frequency bands for training the

LLM later. It assigns each SU $i \in [N]$ to measure the received primary signal power $p_{i,j}^t$ and its noise variance $(\sigma_{i,j}^t)^2$ on each target frequency band $j \in [m_t]$ from his location. Each SU then forms a private Bernoulli belief $\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) \in [0,1]$ as the probability of PU presence (e.g., [26], [27]):

$$\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) = Q_u\left(\sqrt{\frac{2Mp_{i,j}^t}{(\sigma_{i,j}^t)^2}}, \sqrt{\lambda^t}\right),$$

where λ^t is the detection threshold, $Q_u(a,b)$ is the uth-order Marcum Q-function as $Q_u(a,b) = \int_b^\infty x\left(\frac{x}{a}\right)^{u-1}\exp\left(-\frac{x^2+a^2}{2}\right)I_{u-1}(ax)\,dx$ and M is the number of sensing samples. He models PU presence at frequency band j as $p_j^t \sim \text{Bernoulli}(\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t))$, where $p_j^t = 1$ indicates PU presence and $p_j^t = 0$ otherwise.

Each SU i is non-cooperative to have his own selfish objective to receive the maximum credits or rewards from the fusion center. He 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 fusion center. The fusion center and other SUs are uncertain of his $\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)$ realization.

2) Stage II. Online Feedback Aggregation and RLHF Policy Update: After receiving each SU i's sensing 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 fusion center aggregates according to the weight w_i^t for each band $j \in [m_t]$:

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

with a uniform weight w_i^1=1 for all $i \in [N]$ in the first time slot, which is the common RLHF approach using uniform weight for equal SU contribution in the beginning (e.g., [28], [29], [16]). The aggregated sensing outcome $\{\mathcal{P}(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}$ will be included to construct the sensing dataset $\{\mathcal{P}(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}$. Based on this sensing dataset, the system then updates a CSS policy using direct preference optimization (DPO) to solve a KL-regularized optimization problem against the reference policy (e.g., [13]).

3) Reweighing SUs: The current RLHF practices simply average human feedback in the preference aggregation using uniform w_i^t=1 for all $i \in [N]$ and $t \in [T]$ in (1) (e.g., [16]–[18]). To our best knowledge, we are the first to consider dynamic adjustment of each SU's weight in the online RLHF process. The fusion center implements and tests the obtained policy for users' practical usage and learns the realized binary PU presence $p_j^t \in \{1,0\}$ for each frequency band $j \in [m_t]$ according to its users' realized experience. Then, it updates each SU i's weight w_i^{t+1} for next time slot's feedback aggregation:

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

for next slot $t+1$'s aggregation according to sensing 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 realized PU presence $\{p_j^t\}_{j=1}^{m_t}$.

Each SU i is non-cooperative to have his own selfish objective to receive the maximum credits or rewards from the fusion center. He then strategically manipulates his sensing feedback $\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)$ for obtaining a large weight in the fusion center's sensing feedback aggregation in (1) in each time slot $t \in [T]$ and aims to maximize his long-term influence benefit (e.g., [30], [31]) as the cumulative weight over the whole T time slots as follows:

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

On the other hand, the fusion center's PU inference loss based on its aggregation over T time slots is given as follows:

$$L = \sum_{t=1}^T \frac{1}{m_t}\sum_{j=1}^{m_t}\left(\sum_{i=1}^N \frac{w_i^t \hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)}{\sum_{i'=1}^N w_{i'}^t} - p_j^t\right)^2,$$

which is defined as the mean square error (MSE) between the fusion center's weighted aggregation in (1) and the PU presence. It wants to improve the sensing feedback accuracy in the aggregation by assigning the largest weight to the most accurate SU feedback at time t and such assignment will change over time. As each SU senses multiple frequency bands at each time t, the ideal choice for the fusion center is to commit to the SU i^* incurring the least average sensing loss over T time slots:

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

However, the best SU is unknown in the online iteration. The fusion center then turns to reducing the sensing-accuracy regret between online weighted aggregation and offline choice of the best SU in hindsight as follows:

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

Note that the SU utility in (3) may not align with the fusion center's objective in (4), leading to untruthful feedback for a large weight (e.g., [32]). For example, suppose that frequency-band number $m_t = 1$ and time slot number $T = 1$. The fusion center updates $w_i^{t+1} = 1$ if $|\hat{\mathcal{P}}_i^t - p_j^t| \le 0.2$, $w_i^{t+1} = 0.5$ if $|\hat{\mathcal{P}}_i^t - p_j^t| \in (0.2, 0.5]$, and $w_i^{t+1} = 0$ otherwise. An SU i holding $\mathcal{P}_i^t = 0.6$ obtains an expected weight of 0.3 in total by truthfully reporting. However, he can obtain an expected weight of 0.6 by misreporting any $\hat{\mathcal{P}}_i^t \ge 0.8$, increased from being honest. Therefore, it is crucial for the fusion center to properly design the weight update function for any SU's truthful feedback and a small regret.

B. Dynamic Bayesian Game Formulation for RLHF

Based on our system model above, we formulate the multi-agent dynamic CSS between the fusion center and N strategic SUs as a new dynamic Bayesian game in the following:

- In Stage I of each time slot $t \in [T]$, each SU i with his private sensing result $\{\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)\}_{j=1}^{m_t}$ determines

his sensing feedback $\{\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t|x_j^t)\}_{j=1}^{m_t}$ (may not be the truth) to maximize his accumulative weight in (3).

- In Stage III of each time slot $t\in[T]$, the fusion center updates each SU'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 any SU or the fusion center in Stage II. We need to carefully design an online aggregation mechanism for ensuring each SU's truthful sensing feedback and a vanishing regret over time. We define the desired properties as below.

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

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

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. Three Benchmarks: k-out-of-N, Uniform and Median Based RLHF

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

A. Benchmark 1: Model-Based k-out-of-N Scheme

In the CSS literature, the fusion center adopts a k-out-of-N scheme to commit to the sensing feedback that k SUs favor (e.g., [5], [6], [7]). For our CSS problem, the fusion center sets each aggregated sensing outcome $\mathcal{P}(y_{l_j}^t \succ y_{l'_j}^t|x_j^t)=1$ if the number of sensing feedback above 0.5 is no less than k and sets $\mathcal{P}(y_{l_j}^t \succ y_{l'_j}^t|x_j^t)=0$ otherwise. Unfortunately, such an average feedback aggregation scheme can lead to a non-vanishing sensing regret.

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

The proof is given in Appendix A. As the committed SU sensing feedback can be highly noisy, the k-out-of-N scheme fails to identify the most accurate SU in the dynamic CSS process, leading to a non-vanishing time-average regret even if the time slot number $T\to\infty$.

B. Benchmark 2: Average Aggregation for RLHF

The current practice of RLHF simply averages users' feedback in the aggregation using uniform $w_i^t=1$ for all $i\in[N]$ and $t\in[T]$ (e.g., [28], [29], [16]). Unfortunately, such an average feedback aggregation scheme still leads to a non-vanishing sensing regret as shown below.

Lemma 2: The fusion center's sensing regret in (4) under the benchmark 2 of average aggregation 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 proof is given in Appendix B. Benchmark 2 fails to dynamically adjust SUs' aggregation weights according to their online sensing feedback accuracy. Thus, the most accurate SU cannot receive the largest weight in the online learning process, leading to a non-vanishing time-average regret even if the time slot number $T\to\infty$.

C. Benchmark 3: Median Aggregation for RLHF

In the algorithmic game theory literature, the popular "median" scheme is widely used to motivate strategic agents' truthful reporting (e.g., [22], [23]). We define it below.

Definition 3 (Median Aggregation Scheme): The fusion center first re-organizes SUs' sensing 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 frequency band $j\in[m_t]$ in each time slot $t\in[T]$. It then chooses the median $\hat{\mathcal{P}}_{k_s,j}^t$ as its sensing aggregation, where the index $s=\frac{N}{2}$ if N is even and $s=\frac{N+1}{2}$ otherwise.

Since the benchmark 3 independently commits to the median sensing feedback for each frequency band in each time slot, at the equilibrium, all the SUs' sensing feedback will converge to one common point for an equal probability to be the median. We summarize the equilibrium in the following.

Lemma 3: At the equilibrium of the benchmark 3, all the SUs' 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 cannot guarantee to converge to the same as their own sensing results. Accordingly, the fusion center's sensing regret in (4) under the untruthful median scheme is $R_3(T)=\mathcal{O}(T)$, leading to a non-vanishing time-average regret $\lim_{T\to\infty}\frac{R_3(T)}{T}>0$.

The proof is given in Appendix C. The median scheme is not truthful since an SU may be committed with a positive probability by misreporting rather than no probability by truthful sensing feedback, leading to a total accuracy loss of $\mathcal{O}(T)$ over the T time slots. Yet, there can exist an SU o holding $\mathcal{P}_o(y_{l_j}^t \succ y_{l'_j}^t|x_j^t)=p_j^t$, incurring zero accuracy loss of the best fixed SU in hindsight. The average sensing regret is then non-vanishing even if the time slot number $T\to\infty$. Given non-vanishing regrets of all three benchmark schemes, we are well motivated to develop a truthful mechanism to substantially reduce the fusion center's sensing regret.

IV. Our Truthful Online Weighted Aggregation Mechanism

As benchmark 2 of average feedback aggregation fails to achieve a sublinear sensing regret over time, we are well motivated to dynamically adjust each SU's weight according

to his sensing feedback accuracy in each time slot. Unlike benchmark 2, in Stage III of each time slot, we assign a larger weight (compared to the others) if an SU's prior feedback is closer to the realized binary PU presence. We need to carefully design our online mechanism weightage in (2) to ensure that each obtains the largest long-term influence in Definition 1 only with truthful feedback. We define in the following.

Definition 4 (Online Weighted Aggregation Mechanism): At Stage III of each time slot $t \in [T]$, the fusion center updates each SU's weight w_i^{t+1} in (2) based on his sensing feedback $\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)$ and the realized binary PU presence 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 to be determined later.

Intuitively, our mechanism determines each SU's weight in time slot $t+1$ based on his sensing feedback accuracy in the previous time slot t. If the squared difference between his sensing feedback and the realized binary PU presence $(\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t)^2$ is small, his weight w_i^{t+1} will be only reduced by a small value from w_i^t. Though all SUs' weights are decreasing over time, we care about the relative weighted aggregation as in (1) and the SU with a small decrement has a large influence in the fusion center's sensing aggregation. Our mechanism satisfies the truthful property as shown below.

Proposition 1: 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 frequency band $j \in [m_t]$, SU $i \in [N]$ and time slot $t \in [T]$.

The proof is given in Appendix D. As each SU holds a Bernoulli belief on PU presence p_j^t, any deviation from truthful feedback leads to a strictly less weight in any time slot t. Thus, no SU has the incentive to misreport and their truthfulness is guaranteed. Further, our mechanism is efficient and incurs a vanishing time-average regret in T.

Theorem 1: Our online weighted aggregation mechanism in Definition 4 incurs a sublinear regret $R_{\mathcal{M}}(T)$=$\mathcal{O}(T^{\frac{1}{2}})$ by choosing the step size α in (5) as

$$\alpha = \frac{2}{3}\sqrt{\frac{2\ln N}{T}},$$

leading to zero time-average regret with $\lim_{T\to\infty} \frac{R_{\mathcal{M}}(T)}{T} = 0$.

Proof. According to Proposition 1, we have $\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) = \mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t)$ for 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

$$\begin{aligned}
&\ln \frac{\sum_{i=1}^N w_i^{T+1}}{\sum_{i=1}^N w_i^1} \\
=&\ln\left(\sum_{i=1}^N w_i^{T+1}\right) - \ln\left(\sum_{i=1}^N w_i^1\right) \\
=&\ln\left(\sum_{i=1}^N \prod_{t=1}^T (1 - \alpha \frac{1}{m_t} \sum_{j=1}^{m_t} (\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t)^2)\right) - \ln N \\
\geq&\ln\left(\prod_{t=1}^T (1 - \alpha \frac{1}{m_t} \sum_{j=1}^{m_t} (\mathcal{P}_{i^*}(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t)^2)\right) - \ln N \\
=&\sum_{t=1}^T \ln\left(1 - \alpha \frac{1}{m_t} \sum_{j=1}^{m_t} (\mathcal{P}_{i^*}(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t)^2\right) - \ln N \\
\geq& -\alpha \sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} (\mathcal{P}_{i^*}(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t)^2 \\
&- \alpha^2 \sum_{t=1}^T \left(\frac{1}{m_t} \sum_{j=1}^{m_t} (\mathcal{P}_{i^*}(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t)^2\right)^2 - \ln N \\
\geq& -\alpha \sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_{i^*}(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t\right)^2 - \alpha^2 T - \ln N,
\end{aligned} \quad (6)$$

where we choose $\alpha < \frac{1}{2}$ and denote i^* as the best SU in hindsight. The first and the third inequalities hold due to $0 < \alpha < \frac{1}{2}$ and $0 \leq (\mathcal{P}_{i^*}(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t)^2 \leq 1$ 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

$$\begin{aligned}
&\ln \frac{\sum_{i=1}^N w_i^{t+1}}{\sum_{i=1}^N w_i^t} \\
=&\ln\left(\frac{\sum_{i=1}^N w_i^t \cdot (1 - \alpha \frac{1}{m_t} \sum_{j=1}^{m_t} (\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t)^2)}{\sum_{i'=1}^N w_{i'}^t}\right) \\
\leq&\ln\left(\frac{\sum_{i=1}^N w_i^t \cdot e^{-\alpha \frac{1}{m_t} \sum_{j=1}^{m_t} (\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t)^2}}{\sum_{i'=1}^N w_{i'}^t}\right) \\
\leq& -\alpha \frac{1}{m_t} \sum_{j=1}^{m_t} \frac{\sum_{i=1}^N w_i^t (\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t)^2}{\sum_{i'=1}^N w_{i'}^t} + \frac{\alpha^2}{8},
\end{aligned} \quad (7)$$

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 $X = -\frac{1}{m_t} \sum_{j=1}^{m_t} (\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t)^2 \in [-1, 0]$ and $\alpha \in R$, we have

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

According to (7), we have

$$\begin{aligned}
&\ln \frac{\sum_{i=1}^N w_i^{T+1}}{\sum_{i=1}^N w_i^1} \\
=&\ln\left(\frac{\sum_{i=1}^N w_i^{T+1}}{\sum_{i=1}^N w_i^t} \frac{\sum_{i=1}^N w_i^t}{\sum_{i=1}^N w_i^{t-1}} \cdot \cdots \cdot \frac{\sum_{i=1}^N w_i^2}{\sum_{i=1}^N w_i^1}\right) \\
=&\sum_{t=1}^T \ln \frac{\sum_{i=1}^N w_i^{T+1}}{\sum_{i=1}^N w_i^t} \\
\leq& -\alpha \sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} \frac{\sum_{i=1}^N w_i^t (\mathcal{P}_i(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t)^2}{\sum_{i'=1}^N w_{i'}^t} + \frac{\alpha^2 T}{8}.
\end{aligned} \quad (8)$$

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

$$-\alpha \sum_{t=1}^T \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_{i^*}(y_{l_j}^t \succ y_{l'_j}^t | x_j^t) - p_j^t\right)^2 - \alpha^2 T - \ln N$$

$$\leq -\alpha\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}\frac{\sum_{i=1}^{N}w_i^t(\mathcal{P}_i(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)-p_j^t)^2}{\sum_{i'=1}^{N}w_{i'}^t}+\frac{\alpha^2 T}{8}.$$

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

$$\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}\frac{\sum_{i=1}^{N}w_i^t(\mathcal{P}_i(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)-p_j^t)^2}{\sum_{i'=1}^{N}w_{i'}^t}$$
$$-\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}\left(\mathcal{P}_{i^*}(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)-p_j^t\right)^2\leq\frac{\ln N}{\alpha}+\frac{9T\alpha}{8}.$$

Choosing $\alpha=\frac{2}{3}\sqrt{\frac{2\ln N}{T}}<\frac{1}{2}$ (true as $T\to\infty$), we have

$$\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}\frac{\sum_{i=1}^{N}w_i^t(\mathcal{P}_i(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)-p_j^t)^2}{\sum_{i'=1}^{N}w_{i'}^t}$$
$$-\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}\left(\mathcal{P}_{i^*}(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)-p_j^t\right)^2\leq 3\sqrt{\frac{T\ln N}{2}}.$$

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

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

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

According to Theorem 1, our mechanism obviously improves from benchmarks 1-3 by distinguishing the most accurate SU in the dynamic CSS process as $T\to\infty$. As N increases, the fusion center may find a more accurate SU in hindsight. Thus, it chooses a larger step-size α in (5) to punish inaccurate SUs more in the weighted aggregation to retire them. As T increases, the fusion center is more patient in choosing a smaller α in (5) to select the best SU in hindsight with more time slots and samples.

V. Extension to Limited SU Feedback Case in RLHF-Based Cooperative Spectrum Sensing

Recall that in Sections II-IV, we assume that the fusion center has access to all the SUs' sensing feedback per time slot. In practice, waiting for multiple SUs to sense channels and report their measurements can introduce significant delays. Moreover, simultaneous feedback transmissions from multiple SUs may lead to overhead or signal collisions (e.g., [24], [25]). In this section, we extend to consider a challenging case where the fusion center can receive only one SU's report transmission per time slot to ensure fast sensing and immediate action. In the following, we first present our system model for this limited SU feedback scenario and the dynamic Bayesian game formulation. We then give our mechanism design and analysis.

A. System Model of Limited Sensing Feedback from SUs in RLHF-Based Cooperative Spectrum Sensing

Similar to the system model in Section II-A, the fusion center iterates the dynamic CSS process in T time slots, where each time slot $t\in[T]$ still contains three stages. In Stage I, instead of querying all the N SUs' sensing feedback, the fusion center can only select one SU $I_t\in[N]$ in each time slot t to sense the frequency bands for his sensing feedback. We consider that the fusion center uses a mixed strategy to select each SU i with a probability of $\frac{w_i^t}{\sum_{i'\in[N]}w_{i'}^t}$ in each time slot t. In Stage II, after receiving the chosen SU's sensing feedback $\{\hat{\mathcal{P}}_{I_t}(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)\}_{j=1}^{m_t}$, the system determines the sensing outcome for each prompt $j\in[m_t]$ as follows:

$$\mathcal{P}(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)=\hat{\mathcal{P}}_{I_t}(y_{l_j}^t\succ y_{l'_j}^t|x_j^t),\tag{9}$$

which will be included to construct the human sensing dataset $\{\mathcal{P}(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)\}_{j=1}^{m_t}$ for training and updating the RLHF policy later. In Stage III, the fusion center dynamically adjusts each SU's weight in the online RLHF process and determines

$$w_i^{t+1}=f_i(\{\hat{\mathcal{P}}_{I_t}(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)\}_{j=1}^{m_t},\{p_j^t\}_{j=1}^{m_t})\tag{10}$$

for the next slot $t+1$'s selection according to sensing feedback $\{\hat{\mathcal{P}}_{I_t}(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)\}_{j=1}^{m_t}$ and the realized PU presence $\{p_j^t\}_{j=1}^{m_t}$, where $w_i^1=1$ for any $i\in[N]$.

Each SU i is non-cooperative to have his own selfish objective to receive the maximum credits or rewards from the fusion center. He then strategically manipulates his sensing result $\hat{\mathcal{P}}_i(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)$ for obtaining a large weight in the fusion center's sensing feedback aggregation in (1) in each time slot $t\in[T]$, maximizing his long-term influence benefit as the cumulative weights over whole T time slots as in (3).

On the other hand, the fusion center adopts a mixed strategy to choose each SU $i\in[N]$ with a probability of $\frac{w_i^t}{\sum_{i'\in[N]}w_{i'}^t}$ in each time slot t. It wants to improve the sensing feedback accuracy in the aggregation by assigning the largest weight to the most accurate SU. As the best SU is still unknown in the online iteration, it then aims to reduce the sensing regret between online mixed selection and offline choice of the best SU in hindsight, where the PU inference loss is defined as the MSE between the fusion center's mixed selection and the realized binary PU presence as follows:

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

$$:=\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{w_i^t}{\sum_{i'=1}^{N}w_{i'}^t}\hat{\ell}_i^t-\min_{i\in[N]}\sum_{t=1}^{T}\ell_i^t, \tag{11}$$

where we define $\hat{\ell}_i^t:=\frac{1}{m_t}\sum_{j=1}^{m_t}\left(\hat{\mathcal{P}}_i(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)-p_j^t\right)^2$ and $\ell_i^t:=\frac{1}{m_t}\sum_{j=1}^{m_t}\left(\mathcal{P}_i(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)-p_j^t\right)^2$.

B. Dynamic Bayesian Game Formulation for RLHF under Limited SU Feedback

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

- In Stage I of each time slot $t\in[T]$, the fusion center first chooses an SU $I_t\in[N]$ for his sensing feedback. Then, the chosen SU I_t with his private sensing result $\{\mathcal{P}_{I_t}(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)\}_{j=1}^{m_t}$ determines his sensing feedback $\{\hat{\mathcal{P}}_{I_t}(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)\}_{j=1}^{m_t}$ to maximize his accumulative weights in (3).
- In Stage III of each time slot $t\in[T]$, the fusion center updates each SU's weight w_i^{t+1}=$f_i(\{\hat{\mathcal{P}}_{I_t}(y_{l_j}^t\succ y_{l'_j}^t|x_j^t)\}_{j=1}^{m_t},\{p_j^t\}_{j=1}^{m_t})$ for reducing its regret in (11).

Note that there is no strategic decision for any SU or the fusion center in Stage II. We need to carefully design an online mixed selection mechanism for ensuring each SU's truthful sensing feedback and a vanishing regret over time as given in Definitions 1 and 2. Before that, we introduce a heuristic approach based on our mechanism in Definition 4 of the full SU feedback case to check if it still works.

C. Our Truthful Online Mixed Selection Mechanism Design and Analysis

The multi-armed bandit literature (e.g., [33], [34]) commonly introduces an unbiased estimator $\tilde{\ell}_i^t$ of the SU i's feedback loss as follows:

$$\tilde{\ell}_i^t=\begin{cases}\frac{\ell_i^t}{w_i^t/\sum_{i'=1}^{N}w_{i'}^t}, & \text{if } i=I_t,\\ 0, & \text{otherwise,}\end{cases} \tag{12}$$

where the feedback loss of the chosen SU I_t is further divided by his chosen probability to be unbiased and that of each unseen or unchosen SU is set as 0. One may think about if our online weighted aggregation mechanism in Definition 4 is still feasible by simply replacing the feedback loss in (5) with $\tilde{\ell}_i^t$ in (12). Unfortunately, since the estimated loss $\tilde{\ell}_i^t$ are unbounded for $i=I_t$, we cannot ensure that each SU's weight in (5) is positive and bounded in [0, 1]. Thus, we have the following.

Lemma 4: The online weighted aggregation mechanism in Definition 4 is no longer valid in the limited SU feedback case if we simply use the estimated loss $\tilde{\ell}_i^t$ in (12) to replace the feedback loss in (5) for updating each SU i's weight w_i^t, which cannot guarantee that each SU i's weight w_i^t is positive and bounded in [0, 1].

Lemma 4 indicates that it is non-trivial to extend our mechanism from full SU feedback to the limited SU feedback case. As the fusion center can only observe one SU's feedback in each time slot, it needs to carefully balance the exploitation of selected SUs and the exploration of unselected ones during the dynamic CSS process. Further, with such limited feedback information, it is even more challenging to guarantee a sublinear regret. The idea of our new mechanism design is to introduce an exploitation parameter and update the selected SU's weight with some probability. We define our mechanism in the following.

Definition 5 (Online Mixed Selection Mechanism): At Stage III of each time slot $t\in[T]$, the fusion center updates each SU's weight w_i^{t+1} in (10) as follows:

$$w_i^{t+1}=\begin{cases}(1-\beta)\gamma_i^{t+1}+\beta, & \text{if } i=I_t,\\ w_i^t, & \text{otherwise,}\end{cases} \tag{13}$$

where $\beta\in(0,1)$ is an exploitation parameter,

$$\gamma_i^{t+1}=\begin{cases}\gamma_i^t\left(1-\alpha\frac{\hat{\ell}_i^t(1-\alpha/\theta_i^t)}{\theta_i^t}\right), & \text{if } i=I_t,\\ \gamma_i^t, & \text{otherwise,}\end{cases} \tag{14}$$

$\theta_i^t=\frac{w_i^t}{\sum_{i'=1}^{N}w_{i'}^t}$ denotes the probability of selecting each SU, $\gamma_i^1=1$ for $i\in[N]$, and $\alpha\in(0,\theta_i^t)$ is a step-size parameter.

Our mechanism in Definition 5 balances the exploration and the exploitation during the dynamic CSS process. It updates each chosen SU's weight w_i^{t+1} in (13) only with a probability of $1-\beta$ for uniform exploration. Note that if an SU is frequently chosen before, his weight keeps decreasing from 1 and becomes smaller than the others. Therefore, the probability that he will be chosen in future time slots is lower than that of the unchosen SUs, especially when his sensing feedback accuracy is low. Further, a chosen SU i's weight w_i^{t+1} in (13) is proportional to γ_i^{t+1} in (14), which will be decreased by a small value if his feedback loss $\hat{\ell}_i^t$ is small, otherwise large. By carefully designing α and β, our mechanism can guarantee valid w_i^t in (13) and γ_i^t in (14).

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

Proposition 2: Our mechanism in Definition 5 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 frequency band $j\in[m_t]$, SU $i\in[N]$ and time slot $t\in[T]$.

The proof is given in Appendix E. Further, our mechanism is efficient and incurs a vanishing time-average regret in T.

Theorem 2: Our mechanism in Definition 5 incurs the sublinear sensing-accuracy regret $R_{\mathcal{M}}(T)$=$\mathcal{O}(\sqrt{T})$ by choosing step-sizes β in (13) and α in (14) as follows:

$$\beta=2\sqrt{\frac{N\ln N}{7T}},\ \alpha=\sqrt{\frac{\ln N}{7NT}},$$

leading to zero time-average regret with $\lim_{T\to\infty}\frac{R_{\mathcal{M}}(T)}{T}=0$.

The proof is given in Appendix F. According to Theorem 2, our mechanism can distinguish the most accurate SU in the online process as $T\to\infty$. As the SU number N increases, the fusion center faces with more uncertainty of feedback accuracy in exploration than exploitation. Thus, it becomes more patient with a larger exploitation parameter β and a smaller step-size α to punish inaccurate (chosen) SUs less in the weight update. On the other hand, as the time slot number T increases, the fusion center has more room to explore for the most accurate SU with more time slots and samples. Thus, it chooses a smaller exploitation parameter β and a smaller step-size α to punish inaccurate (chosen) SUs less in the weight update.

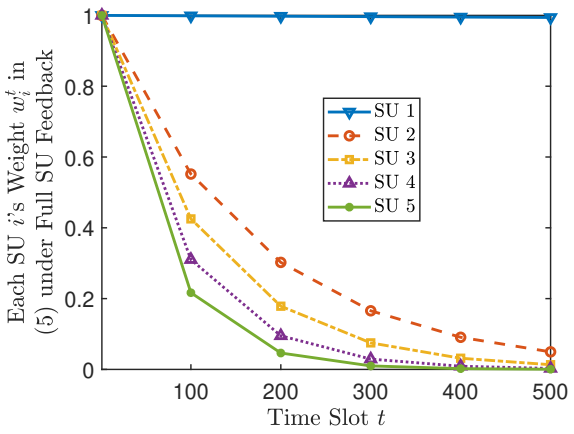

Fig. 2. Each SU i's weight w_i^t versus time slot t under full feedback. Here we fix SU number $N = 5$, total time slot number $T = 500$, and prompt number $m_t = 20$.

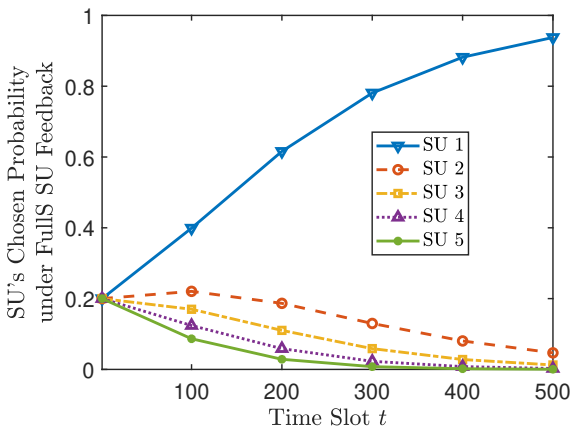

Fig. 3. Each SU i's chosen probability $\frac{w_i^t}{\sum_{i'=1}^N w_{i'}^t}$ versus time slot t under full feedback. Here we fix SU number $N = 5$, total time slot number $T = 500$, and prompt number $m_t = 20$.

VI. Simulation Experiments

In this section, we run simulations to show our mechanism's great improvement over the three benchmark schemes. The fusion center chooses 20 frequency bands and construct $m_t = 20$ prompts to query in each time slot $t \in [T]$, where the channel states are unknown to the fusion center. After obtaining SUs' sensing feedback on all 20 frequency bands in each time slot t, the fusion center uses direct preference optimization (DPO) for the RLHF policy training and update (e.g., [16]). Following the RLHF literature (e.g., [35], [30]), we evaluate the fusion center's sensing accuracy on feedback aggregation, which directly impacts the performance of the final fine-tuned CSS policy.

For ease of exposition and illustration, we first consider $N = 5$ SUs and index SUs in a decreasing order of sensing result accuracy, i.e., SU 1 to be the most accurate and SU 5 to be the least. To further test whether our mechanisms are robust to distinguish the most accurate SU, we consider SU 1 as the perfect SU with small sensing errors and the remaining 4 SUs with larger sensing errors. We use synthetic data to randomly generate each SU i's sensing result $\mathcal{P}_i(y_{l_j}^t \succ y_{l_j'}^t | x_j^t)$ in the range of [0, 1]. Further, we randomly generate the binary realized PU presence $p_j^t \in \{1, 0\}$ for each frequency band $j \in [m_t]$ and fix prompt number m_t=20.

Figure 2 shows how each SU i's weight w_i^t evolves over time slot $t \in [T]$ under full SU feedback, where each SU's sensing feedback is available to the fusion center in each time slot t. As t increases to 500, our mechanism manages to assign the largest weight to the most accurate SU 1 and assign much smaller weights to the remaining ones especially for the most inaccurate SU 5, which is consistent with each SU's weight (5) in Definition 4.

Figure 3 is similar to Fig. 2 and shows how each SU i's chosen probability $\frac{w_i^t}{\sum_{i'=1}^N w_{i'}^t}$ evolves over time slot t. As t increases to 500, our mechanism manages to assign the largest probability to the most accurate SU 1 (over 0.9) and assign near-zero probabilities to the remaining ones especially for the most inaccurate SU 5, which further verifies the effectiveness of our mechanism in Definition 4.

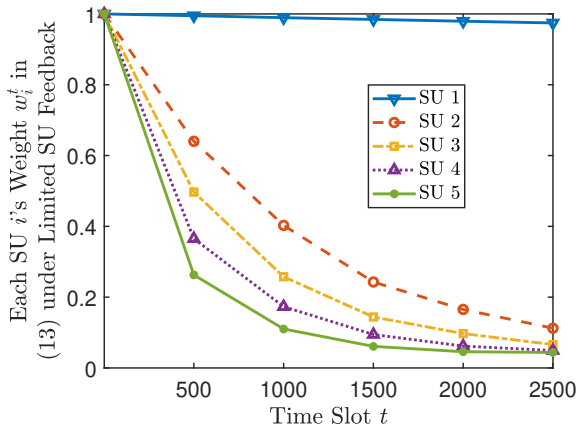

Fig. 4. Each SU i's weight w_i^t versus time slot t under limited SU feedback. Here we fix SU number $N = 5$, total time slot number $T = 2500$, and prompt number $m_t = 20$.

Figure 4 shows how each SU i's weight w_i^t evolves over time slot $t \in [T]$ under limited SU feedback, where only one selected SU's sensing feedback is available to the fusion center in each time slot t. It indicates that our mechanism still manages to assign the largest weight to the most accurate SU 1 overtime, which is consistent with (13) in Definition 5. Yet, as the fusion center only has access to one selected SU's feedback in each time slot, our mechanism needs more time slots to assign relatively small weights to the remaining ones compared to Fig. 2 under full feedback.

Figure 5 is similar to Fig. 4 and shows how each SU i's chosen probability $\frac{w_i^t}{\sum_{i'=1}^N w_{i'}^t}$ evolves over time slot $t \in [T]$ under limited feedback. It indicates that our mechanism manages to assign the largest probability to the most accurate SU 1 (near 0.8), which further verifies the effectiveness of our mechanism in Definition 5. Yet, under limited feedback, our mechanism needs more time slots (as t increases to 2500) to converge to

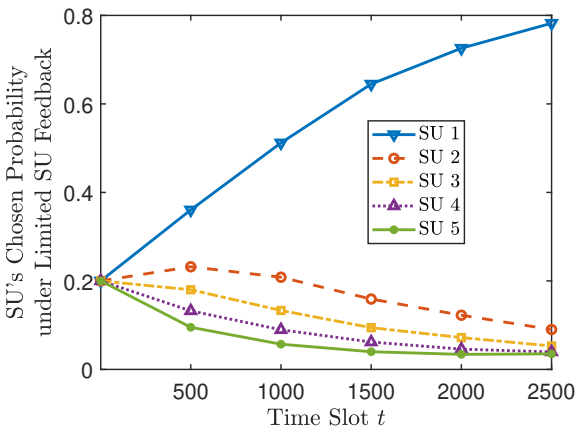

Fig. 5. Each SU i's chosen probability $\frac{w_i^t}{\sum_{i'=1}^{N} w_{i'}^t}$ versus time slot t under limited SU feedback. Here we fix SU number $N = 5$, total time slot number $T = 2500$, and prompt number $m_t = 20$.

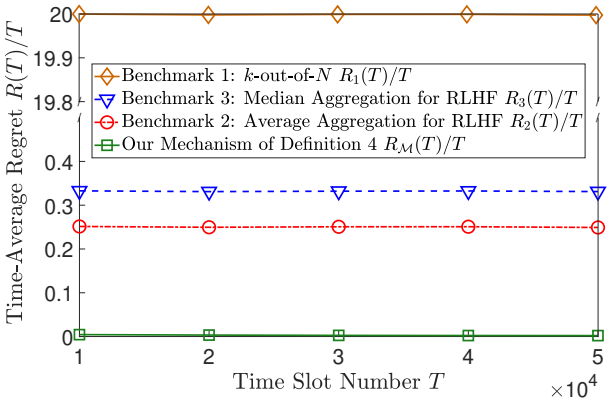

Fig. 6. Time-average regrets of benchmarks 1, 2, 3 and our mechanism under full SU feedback versus the time slot number T, respectively. Here we choose $k = N/2$ for benchmark 1 of the k-out-of-N scheme, consider a large SU scale of $N = 50$ and fix prompt number $m_t = 20$.

assign relatively small probabilities to the remaining ones.

Figure 6 shows the time-average regrets $R(T)/T$ of benchmarks 1, 2, 3, and our mechanism under full SU feedback versus the time slot number T under a large SU scale of $N = 50$. We find that the fusion center's time-average regret is greatly reduced by our mechanism from the three benchmarks. Besides, time-average regrets of benchmarks 1-3 do not decrease with T and are always greater than zero, respectively, consistent with Lemmas 1, 2 and 3. Differently, our mechanism's time-average regret decreases with T and tends to 0, consistent with Theorem 1.

Figure 7 shows the time-average regrets $R(T)/T$ of benchmarks 1, 2, 3, and our mechanism under limited feedback versus the time slot number T under a large SU scale of $N = 50$. We find that the fusion center's time-average regret is still greatly reduced by our mechanism from the three benchmarks. Besides, time-average regrets of benchmarks 1-3 do not decrease with T and are always greater than zero, respectively, consistent with Lemmas 1, 2 and 3. Note that compared with the full feedback case in Fig. 6, our mechanism

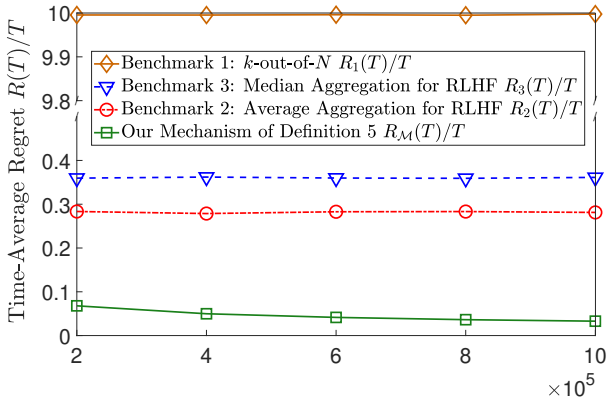

Fig. 7. Time-average regrets of benchmarks 1, 2, 3 and our mechanism under limited SU feedback versus the time slot number T, respectively. Here we choose $k = 5$ for benchmark 1 of the k-out-of-N scheme, consider a large SU scale of $N = 50$ and fix the prompt number $m_t = 20$. For a fair comparison, we randomly select 10 SUs' feedback out of the total 50 for the three benchmarks to make decision.

needs more time slots to obtain small enough time-average regret under limited feedback. Nevertheless, our mechanism's time-average regret still decreases with T and tends to 0, consistent with Theorem 2.

VII. Conclusion

In this paper, we are the first to leverage RLHF for dynamic CSS. We first show that existing CSS approaches (i.e., k-out-of-N and direct application of RLHF with uniform SU weighting) incur a non-vanishing regret of $\mathcal{O}(T)$ over a time horizon T. To address this, we design an online weighted aggregation mechanism within the RLHF framework that dynamically adjusts each SU's weight based on historical feedback. This approach not only incentivizes truthful reporting from SUs but also achieves a sublinear regret $\mathcal{O}(\sqrt{T})$ over time. Furthermore, we extend our RLHF design to settings with limited SU feedback per time slot. Simulation results validate significant performance gains of our proposed mechanisms compared to benchmark schemes.

One future direction is to consider the case that an SU can only sense a subset of frequency bands. In this way, the fusion center is restricted to assign SUs of a limited set for spectrum sensing and it is non-trivial to extend our analysis under both full and limited SU feedback cases.

References

[1] S. Hao and L. Duan, "Online learning from strategic human feedback in llm fine-tuning," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[2] C.-X. Wang, X. You, X. Gao, X. Zhu, Z. Li, C. Zhang, H. Wang, Y. Huang, Y. Chen, H. Haas *et al.*, "On the road to 6g: Visions, requirements, key technologies, and testbeds," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 905–974, 2023.

[3] G. Liu, R. Xi, Z. Han, L. Han, X. Zhang, L. Ma, Y. Wang, M. Lou, J. Jin, Q. Wang *et al.*, "Cooperative sensing for 6g mobile cellular networks: feasibility, performance and field trial," *IEEE Journal on Selected Areas in Communications*, 2024.

[4] L. Duan, A. W. Min, J. Huang, and K. G. Shin, "Attack prevention for collaborative spectrum sensing in cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 9, pp. 1658–1665, 2012.

[5] K. Cichoń, A. Kliks, and H. Bogucka, "Energy-efficient cooperative spectrum sensing: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1861–1886, 2016.

[6] B. Aygun and A. M. Wyglinski, "A voting-based distributed cooperative spectrum sensing strategy for connected vehicles," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5109–5121, 2016.

[7] D. Janu, K. Singh, and S. Kumar, "Machine learning for cooperative spectrum sensing and sharing: A survey," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 1, p. e4352, 2022.

[8] F. Jiang, L. Dong, Y. Peng, K. Wang, K. Yang, C. Pan, and X. You, "Large ai model empowered multimodal semantic communications," *IEEE Communications Magazine*, 2024.

[9] H. Zhou, C. Hu, D. Yuan, Y. Yuan, D. Wu, X. Liu, and C. Zhang, "Large language model (llm)-enabled in-context learning for wireless network optimization: A case study of power control," *arXiv preprint arXiv:2408.00214*, 2024.

[10] W. Lee and J. Park, "Llm-empowered resource allocation in wireless communications systems," *arXiv preprint arXiv:2408.02944*, 2024.

[11] 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.

[12] 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.

[13] 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.

[14] W. Yuan, H. Leung, S. Chen, and W. Cheng, "A distributed sensor selection mechanism for cooperative spectrum sensing," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6033–6044, 2011.

[15] Z. Jiang, W. Yuan, H. Leung, X. You, and Q. Zheng, "Coalition formation and spectrum sharing of cooperative spectrum sensing participants," *IEEE transactions on cybernetics*, vol. 47, no. 5, pp. 1133–1146, 2016.

[16] 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.

[17] 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.

[18] 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.

[19] B. Wang, K. R. Liu, and T. C. Clancy, "Evolutionary cooperative spectrum sensing game: how to collaborate?" *IEEE transactions on communications*, vol. 58, no. 3, pp. 890–900, 2010.

[20] W. Wang, H. Li, Y. Sun, and Z. Han, "Securing collaborative spectrum sensing against untrustworthy secondary users in cognitive radio networks," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 1–15, 2009.

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

[22] 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.

[23] 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.

[24] Y. Liu, Y. Deng, A. Nallanathan, and J. Yuan, "Machine learning for 6g enhanced ultra-reliable and low-latency services," *IEEE Wireless Communications*, vol. 30, no. 2, pp. 48–54, 2023.

[25] E. J. T. Pereira, D. A. Guimarães, and R. Shrestha, "Vlsi architectures and hardware implementation of ultra low-latency and area-efficient pietra-ricci index detector for spectrum sensing," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2024.

[26] F. F. Digham, M.-S. Alouini, and M. K. Simon, "On the energy detection of unknown signals over fading channels," in *IEEE International Conference on Communications, 2003. ICC'03.*, vol. 5. Ieee, 2003, pp. 3575–3579.

[27] S. Atapattu, C. Tellambura, and H. Jiang, "Energy detection based cooperative spectrum sensing in cognitive radio networks," *IEEE Transactions on wireless communications*, vol. 10, no. 4, pp. 1232–1241, 2011.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] W. Feng, X. Gao, P. Zhao, and S. C. Hoi, "A unified framework for bandit online multiclass prediction," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[34] Y. Huang, Q. Liu, and J. Xu, "Adversarial combinatorial bandits with switching cost and arm selection constraints," in *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*. IEEE, 2024, pp. 371–380.

[35] 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.

Shugang Hao (M'22) received the Ph.D. degree from Singapore University of Technology and Design (SUTD) in 2022. He is a postdoctoral research fellow at SUTD from Sep. 2022. His research interests are RLHF, game theory and mechanism design. He served as the web chair of ACM SenSys 2024, the local arrangement chair of AIoTSys 2024 and the local arrangement chair of IEEE WiOpt 2023.

Lingjie Duan (S'09-M'12-SM'17) received the Ph.D. degree from The Chinese University of Hong Kong in 2012. He is an Associate Professor and Associate Head of Pillar of Engineering Systems and Design with the Singapore University of Technology and Design (SUTD). In 2011, he was a Visiting Scholar at University of California at Berkeley, Berkeley, CA, USA. His research interests include network economics and game theory, cognitive communications, and cooperative networking. He is an Associate Editor of IEEE Transactions on Mobile Computing and IEEE Transactions on Networking. He was an Editor of IEEE Transactions on Wireless Communications and IEEE Communications Surveys and Tutorials. He also served as a Guest Editor of the IEEE Journal on Selected Areas in Communications Special Issue on Human-in-the-Loop Mobile Networks, as well as IEEE Wireless Communications Magazine. He received the SUTD Excellence in Research Award in 2016 and the 10th IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2015. He served as the general chair of IEEE WiOpt 2023.

APPENDIX A
PROOF OF LEMMA 1

We want to prove $R_1(T)=\mathcal{O}(T)$ with a possible sequence of SUs' preferences. In particular, we consider $\mathcal{P}_o(y_{l_j}^t \succ y_{l_j'}^t|x_j^t) = p_j^t$ holds for one particular $o \in [N]$ with any $j \in [m_t]$ and $t \in [T]$. For the remaining SUs, we consider at least k SUs hold sensing results greater than 0.5 when the actual PU presence is 0 and less than k SUs hold sensing results greater than 0.5 when the actual PU presence is 1 in each time slot $t \in [T]$, leading to $\hat{\mathcal{P}}(y_{l_j}^t \succ y_{l_j'}^t|x_j^t) = 1-p_j^t$. Accordingly, we have the best-fixed SU in hindsight is $i^*=o$, which brings

$$\min_{i\in[N]} \sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_i(y_{l_j}^t \succ y_{l_j'}^t|x_j^t) - p_j^t\right)^2 = 0.$$

However, with the fusion center's k-out-of-N scheme, we have the cumulative aggregation loss over T slots as follows:

$$\begin{aligned}&\sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\hat{\mathcal{P}}(y_{l_j}^t \succ y_{l_j'}^t|x_j^t) - p_j^t\right)^2\\ =&\sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(1 - p_j^t - p_j^t\right)^2 = T,\end{aligned}$$

where the last equality holds for either $p_j^t=1$ or 0. Finally, we have the regret for benchmark 1 of k-out-of-N as follows:

$$\begin{aligned}R_1(T) =&\sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\hat{\mathcal{P}}(y_{l_j}^t \succ y_{l_j'}^t|x_j^t) - p_j^t\right)^2\\ &- \min_{i\in[N]} \sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_i(y_{l_j}^t \succ y_{l_j'}^t|x_j^t) - p_j^t\right)^2\\ =&T = \mathcal{O}(T),\end{aligned}$$

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

APPENDIX B
PROOF OF LEMMA 2

We want to prove $R_2(T)=\mathcal{O}(T)$ with a possible sequence of SUs' preferences. In particular, we consider $\mathcal{P}_o(y_{l_j}^t \succ y_{l_j'}^t|x_j^t) = p_j^t$ holds for one particular $o \in [N]$ with any $j \in [m_t]$ and $t \in [T]$. For the remaining SUs, we consider $(\sum_{i=1,i\neq o}^{N} \frac{1}{N}\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l_j'}^t|x_j^t) - \frac{N-1}{N}p_j^t)^2 = c_j^t$ for each $i \neq o, i \in [N]$, $j \in [m_t]$ and $t \in [T]$, where $c_j^t \in [\frac{1}{2}, 1]$. Accordingly, we have the best-fixed SU in hindsight is $i^*=o$, which brings

$$\min_{i\in[N]} \sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_i(y_{l_j}^t \succ y_{l_j'}^t|x_j^t) - p_j^t\right)^2 = 0.$$

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

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

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 for benchmark 2 of average aggregation as follows:

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

which indicates that $\lim_{T\to\infty} \frac{R_2(T)}{T} > 0$. We finish the proof.

APPENDIX C
PROOF OF LEMMA 3

First, we want to prove that $\hat{\mathcal{P}}_i^*(y_{l_j}^t \succ y_{l_j'}^t|x_j^t) = \hat{\mathcal{P}}(y_{l_j}^t \succ y_{l_j'}^t|x_j^t) \in [0,1]$ for $j \in [m_t]$, $i \in [N]$ and $t \in [T]$ is an equilibrium. Note that the fusion center independently determines the aggregation as the median feedback for each prompt $j \in [m_t]$ in each time slot $t \in [T]$, each SU $i \in [N]$ aims to maximize each w_i^t to obtain a largest possible accumulative weight overtime. Since the fusion center always commits to the median feedback, given the other $N-1$ SUs except for i choose $\hat{\mathcal{P}}_k(y_{l_j}^t \succ y_{l_j'}^t|x_j^t)$=$\hat{\mathcal{P}}(y_{l_j}^t \succ y_{l_j'}^t|x_j^t)$, $k \neq i$, $k \in [N]$, SU i feedback any $\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l_j'}^t|x_j^t) \neq \hat{\mathcal{P}}(y_{l_j}^t \succ y_{l_j'}^t|x_j^t)$ will lead to his weight $w_i^t = 0$ because such feedback cannot be the median. Thus, he will feedback consistently as $\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l_j'}^t|x_j^t)$=$\hat{\mathcal{P}}(y_{l_j}^t \succ y_{l_j'}^t|x_j^t)$ for an equal chance to be the median and will never deviate from this feedback strategy.

Then, we want to prove $R_3(T)=\mathcal{O}(T)$ with a possible sequence of SUs' preferences. In particular, we consider $\mathcal{P}_o(y_{l_j}^t \succ y_{l_j'}^t|x_j^t) = p_j^t$ holds for one particular $o \in [N]$ with any $j \in [m_t]$ and $t \in [T]$. Further, we consider $(\hat{\mathcal{P}}_{j,k_m}^t - p_j^t)^2 = c_j^t$ for $j \in [m_t]$ and $t \in [T]$, where $\hat{\mathcal{P}}_{j,k_m}^t$ denotes the median of SUs' feedback $\{\hat{\mathcal{P}}_i(y_{l_j}^t \succ y_{l_j'}^t|x_j^t)\}_{i=1}^N$ and $c_j^t \in [\frac{1}{2}, 1]$. Accordingly, we have the best-fixed SU in hindsight is $i^*=o$, which brings

$$\min_{i\in[N]} \sum_{t=1}^{T} \frac{1}{m_t} \sum_{j=1}^{m_t} \left(\mathcal{P}_i(y_{l_j}^t \succ y_{l_j'}^t|x_j^t) - p_j^t\right)^2 = 0.$$

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

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

$$=\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}\left(\hat{\mathcal{P}}^t_{j,k_m}-y^t_j\right)^2$$
$$=\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}c^t_j=\mathcal{O}(T),$$

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 of the median scheme as follows:

$$R_3(T)=\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}\left(\sum_{i=1}^{N}\frac{w^t_i\hat{\mathcal{P}}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)}{\sum_{i'=1}^{N}w^t_{i'}}-p^t_j\right)^2$$
$$-\min_{i\in[N]}\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j=1}^{m_t}\left(\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)-p^t_j\right)^2$$
$$=\mathcal{O}(T).$$

We then finish the proof.

Appendix D
Proof of Proposition 1

According to our system model in Section II, each SU believes that $p^t_j\sim\texttt{Bernoulli}(\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j))$, we have expectation on w^{t+1}_i in (5) over p^t_j is

$$\mathbb{E}[w^{t+1}_i]$$
$$=w^t_i\frac{1}{m_t}\sum_{j=1}^{m_t}\Big[1-\alpha\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)(\hat{\mathcal{P}}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)-1)^2$$
$$-\alpha(1-\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j))(\hat{\mathcal{P}}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)-0)^2\Big]$$
$$=w^t_i\frac{1}{m_t}\sum_{j=1}^{m_t}\Big[1-\alpha(\hat{\mathcal{P}}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)-\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j))^2$$
$$-\alpha(\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)-\mathcal{P}^2_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j))\Big],$$

which is maximized at $\hat{\mathcal{P}}^*_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)=\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)$. To obtain the largest possible accumulative weight, each SU will truthfully feedback his sensing result 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 finish the proof.

Appendix E
Proof of Proposition 2

Accoridng to our system model in Section V, each SU believes that $p^t_j\sim\texttt{Bernoulli}(\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j))$ and $Pr(I_t=i)=\theta^t_i$, we have expectation on w^{t+1}_i in (13) over p^t_j and I_t is

$$\mathbb{E}[w^{t+1}_i]$$
$$=(1-\beta)\gamma^t_i\frac{1}{m_t}\sum_{j=1}^{m_t}\Big[1-\alpha\Big(1-\frac{\alpha}{\theta^t_i}\Big)\Big(\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)$$
$$(\hat{\mathcal{P}}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)-1)^2+(1-\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j))$$
$$(\hat{\mathcal{P}}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)-0)^2\Big)\Big]+\beta$$
$$=(1-\beta)\gamma^t_i\frac{1}{m_t}\sum_{j=1}^{m_t}\Big[1-\alpha\Big(1-\frac{\alpha}{\theta^t_i}\Big)\Big((\hat{\mathcal{P}}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)$$
$$-\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j))^2+\Big(\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)$$
$$-\mathcal{P}^2_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)\Big)\Big)\Big]+\beta,$$

which is maximized at $\hat{\mathcal{P}}^*_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)=\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)$ if $\alpha<\theta^t_i$. According to the choices of α and β in Theorem 2, we have $\theta^t_i\geq\frac{\beta}{N}=2\alpha>\alpha$, implying $\hat{\mathcal{P}}^*_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)=\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)$. To obtain the largest possible accumulative weight, each SU will truthfully feedback his sensing result 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.

Appendix F
Proof of Theorem 2

According to Proposition 2, we have $\hat{\mathcal{P}}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)=\mathcal{P}_i(y^t_{l_j}\succ y^t_{l'_j}|x^t_j)$ for all $j\in[m_t]$, $i\in[N]$ and $t\in[T]$. To derive a lower-bound on $\ln\frac{\sum_{i=1}^{N}\gamma^{T+1}_i}{\sum_{i=1}^{N}\gamma^1_i}$, we assume $\alpha N\leq\frac{\beta}{2}$ and have

$$\ln\frac{\sum_{i=1}^{N}\gamma^{T+1}_i}{\sum_{i=1}^{N}\gamma^1_i}=\ln\left(\sum_{i=1}^{N}\gamma^{T+1}_i\right)-\ln\left(\sum_{i=1}^{N}\gamma^1_i\right)$$
$$=\ln\left(\sum_{i=1}^{N}\prod_{t=1}^{T}(1-\alpha\tilde{\ell}^t_i)\right)-\ln N$$
$$\geq\ln\left(\prod_{t=1}^{T}(1-\alpha\tilde{\ell}^t_{i^*})\right)-\ln N$$
$$=\sum_{t=1}^{T}\ln\left(1-\alpha\tilde{\ell}^t_{i^*}\right)-\ln N$$
$$\geq-\alpha\sum_{t=1}^{T}\tilde{\ell}^t_{i^*}-\alpha^2\sum_{t=1}^{T}\left(\tilde{\ell}^t_{i^*}\right)^2-\ln N,\quad(15)$$

where we choose

$$\tilde{\ell}^t_i=\begin{cases}\frac{\ell^t_i(1-\alpha/\theta^t_i)}{\theta^t_i}, & \text{if } i=I_t,\\ 0, & \text{otherwise,}\end{cases}$$

$\alpha\tilde{\ell}^t_i\leq\alpha\frac{\ell^t_i(1-\alpha/\theta^t_i)}{\theta^t_i}\leq\alpha\frac{1}{\theta^t_i}\leq\alpha\frac{N}{\beta}\leq\frac{1}{2}$ and denote i^* as the best SU in hindsight. Note that $\theta^t_i\geq\frac{\beta}{N}$ is equal to $(\gamma^t_i+\beta)N\geq\beta\sum_{i=1}^{N}\gamma^t_i$, which holds due to $\gamma^t_i+\beta>\beta>0$ and $N>\sum_{i=1}^{N}\gamma^t_i$. The first inequality holds due to $\alpha\tilde{\ell}^t_i\leq\frac{1}{2}$ 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 $\ln\frac{\sum_{i=1}^{N}\gamma^{t+1}_i}{\sum_{i=1}^{N}\gamma^t_i}$, we have

$$\ln\frac{\sum_{i=1}^{N}\gamma^{t+1}_i}{\sum_{i=1}^{N}\gamma^t_i}=\ln\left(\frac{\sum_{i=1}^{N}\gamma^t_i\cdot(1-\alpha\tilde{\ell}^t_i)}{\sum_{i'=1}^{N}\gamma^t_{i'}}\right)$$
$$=\ln\left(1-\alpha\frac{\sum_{i=1}^{N}\gamma^t_i\cdot\tilde{\ell}^t_i}{\sum_{i'=1}^{N}\gamma^t_{i'}}\right)$$

$$\leq -\alpha\frac{\sum_{i=1}^N \gamma_i^t\cdot\tilde{\ell}_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}+\frac{1}{2}\alpha^2\left(\frac{\sum_{i=1}^N\gamma_i^t\cdot\tilde{\ell}_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}\right)^2$$
$$\leq -\alpha\frac{\sum_{i=1}^N \gamma_i^t\cdot\tilde{\ell}_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}+\frac{1}{2}\alpha^2\frac{\sum_{i=1}^N\gamma_i^t\cdot(\tilde{\ell}_i^t)^2}{\sum_{i'=1}^N\gamma_{i'}^t}, \tag{16}$$

where the first inequality holds due to $\ln(1-\alpha x)\leq -\alpha x+\frac{1}{2}\alpha^2x^2$ for $x=\frac{\sum_{i=1}^N\gamma_i^t\cdot\tilde{\ell}_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}$ and $\alpha x\leq 1/2$. The second inequality holds due to Jensen's inequality. According to (16), we have

$$\begin{aligned}
&\ln\frac{\sum_{i=1}^N\gamma_i^{T+1}}{\sum_{i=1}^N\gamma_i^1}\\
=&\ln\left(\frac{\sum_{i=1}^N\gamma_i^{T+1}}{\sum_{i=1}^N\gamma_i^t}\frac{\sum_{i=1}^N\gamma_i^t}{\sum_{i=1}^N\gamma_i^{t-1}}\cdot\cdots\cdot\frac{\sum_{i=1}^N\gamma_i^2}{\sum_{i=1}^N\gamma_i^1}\right)\\
=&\sum_{t=1}^T\ln\frac{\sum_{i=1}^N\gamma_i^{T+1}}{\sum_{i=1}^N\gamma_i^t}\\
\leq&-\alpha\sum_{t=1}^T\frac{\sum_{i=1}^N\gamma_i^t\cdot\tilde{\ell}_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}+\frac{1}{2}\alpha^2\sum_{t=1}^T\frac{\sum_{i=1}^N\gamma_i^t\cdot(\tilde{\ell}_i^t)^2}{\sum_{i'=1}^N\gamma_{i'}^t}.
\end{aligned}\tag{17}$$

According to (15) and (17), we have

$$\begin{aligned}
&-\alpha\sum_{t=1}^T\tilde{\ell}_{i^*}^t-\alpha^2\sum_{t=1}^T\left(\tilde{\ell}_{i^*}^t\right)^2-\ln N\\
\leq&-\alpha\sum_{t=1}^T\frac{\sum_{i=1}^N\gamma_i^t\cdot\tilde{\ell}_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}+\frac{1}{2}\alpha^2\sum_{t=1}^T\frac{\sum_{i=1}^N\gamma_i^t\cdot(\tilde{\ell}_i^t)^2}{\sum_{i'=1}^N\gamma_{i'}^t}.
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{t=1}^T\frac{\sum_{i=1}^N\gamma_i^t\cdot\tilde{\ell}_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}-\sum_{t=1}^T\tilde{\ell}_{i^*}^t\\
\leq&\frac{\ln N}{\alpha}+\frac{1}{2}\alpha\sum_{t=1}^T\frac{\sum_{i=1}^N\gamma_i^t\cdot(\tilde{\ell}_i^t)^2}{\sum_{i'=1}^N\gamma_{i'}^t}+\alpha\sum_{t=1}^T\left(\tilde{\ell}_{i^*}^t\right)^2.
\end{aligned}$$

After taking expectation of $\tilde{\ell}_i^t$ on the above inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T\frac{\sum_{i=1}^N\gamma_i^t\cdot\tilde{\ell}_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}-\sum_{t=1}^T\tilde{\ell}_{i^*}^t\right]\\
\leq&\frac{\ln N}{\alpha}+\frac{1}{2}\alpha\sum_{t=1}^T\sum_{i=1}^N\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}\frac{(\ell_i^t)^2}{\theta_i^t}+\alpha\sum_{t=1}^T\frac{(\ell_{i^*}^t)^2}{\theta_i^t}\\
\leq&\frac{\ln N}{\alpha}+\frac{1}{2}\alpha NT+\alpha\sum_{t=1}^T\frac{(\ell_{i^*}^t)^2}{\theta_i^t},
\end{aligned}\tag{18}$$

where the second inequality holds due to $\ell_i^t\in[0,1]$ for $i\in[N]$, $\theta_i^t\geq\min\{\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t},\frac{1}{N}\}$ and $\sum_{t=1}^T\sum_{i=1}^N\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}\frac{1}{\theta_i^t}\leq\sum_{t=1}^T\sum_{i=1}^N\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}\frac{1}{\min\{\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t},\frac{1}{N}\}}=NT$. We then derive a lower bound of the expectation of $\sum_{t=1}^T\frac{\sum_{i=1}^N\gamma_i^t\cdot\tilde{\ell}_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}-\sum_{t=1}^T\tilde{\ell}_{i^*}^t$. After taking expectation of $\hat{\ell}_i^t$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T\frac{\sum_{i=1}^N\gamma_i^t\cdot\tilde{\ell}_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}-\sum_{t=1}^T\tilde{\ell}_{i^*}^t\right]\\
=&\sum_{t=1}^T\sum_{i=1}^N\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}\ell_i^t\left(1-\frac{\alpha}{\theta_i^t}\right)-\sum_{t=1}^T\ell_{i^*}^t\left(1-\frac{\alpha}{\theta_i^t}\right)\\
=&\sum_{t=1}^T\sum_{i=1}^N\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}\ell_i^t-\sum_{t=1}^T\ell_{i^*}^t-\sum_{t=1}^T\sum_{i=1}^N\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}\frac{\alpha\ell_i^t}{\theta_i^t}\\
&+\sum_{t=1}^T\ell_{i^*}^t\frac{\alpha}{\theta_i^t}\\
\geq&\sum_{t=1}^T\sum_{i=1}^N\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}\ell_i^t-\sum_{t=1}^T\ell_{i^*}^t-\alpha NT+\alpha\sum_{t=1}^T\frac{(\ell_{i^*}^t)^2}{\theta_i^t},
\end{aligned}\tag{19}$$

where the first equality holds due to $\ell_i^t\in[0,1]$ for $i\in[N]$, $\theta_i^t\geq\min\{\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t},\frac{1}{N}\}$ and $\sum_{t=1}^T\sum_{i=1}^N\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}\frac{1}{\theta_i^t}\leq\sum_{t=1}^T\sum_{i=1}^N\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}\frac{1}{\min\{\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t},\frac{1}{N}\}}=NT$. According to (18) and (19), we have

$$\sum_{t=1}^T\sum_{i=1}^N\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}\ell_i^t-\sum_{t=1}^T\ell_{i^*}^t\leq\frac{3}{2}\alpha NT+\frac{\ln N}{\alpha}.$$

Since $\theta_i^t=\frac{(1-\beta)\gamma_i^t+\beta}{(1-\beta)\sum_{i=1}^N\gamma_i^t+\beta N}<\frac{\gamma_i^t}{\sum_{i=1}^N\gamma_i^t}+\frac{1}{N}$, we further have

$$\begin{aligned}
&\sum_{t=1}^T\sum_{i=1}^N\left(\theta_i^t-\frac{1}{N}\right)\ell_i^t-\sum_{t=1}^T\ell_{i^*}^t\\
<&\sum_{t=1}^T\sum_{i=1}^N\frac{\gamma_i^t}{\sum_{i'=1}^N\gamma_{i'}^t}\ell_i^t-\sum_{t=1}^T\ell_{i^*}^t\\
\leq&3\alpha NT+\frac{\ln N}{\alpha},
\end{aligned}$$

which is equal to

$$\begin{aligned}
\sum_{t=1}^T\sum_{i=1}^N\theta_i^t\ell_i^t-\sum_{t=1}^T\ell_{i^*}^t&\leq\sum_{i=1}^N\sum_{t=1}^T\frac{1}{N}\ell_i^t+3\alpha NT+\frac{\ln N}{\alpha}\\
&\leq\sum_{i=1}^N\sum_{t=1}^T\frac{2\beta}{N}\ell_i^t+3\alpha NT+\frac{\ln N}{\alpha}\\
&\leq 2\beta T+3\alpha NT+\frac{\ln N}{\alpha},
\end{aligned}$$

where the second inequality holds due to $\beta\leq\frac{1}{2}$ and the third due to $\ell_i^t\leq 1$. By taking $\beta=2\alpha N$, we have

$$\begin{aligned}
R_{\mathcal{M}}(T)=\sum_{t=1}^T\sum_{i=1}^N\theta_i^t\ell_i^t-\sum_{t=1}^T\ell_{i^*}^t&\leq 7\alpha NT+\frac{\ln N}{\alpha}\\
&\leq 2\sqrt{7}\sqrt{NT\ln N}=\mathcal{O}(\sqrt{T})
\end{aligned}$$

at $\alpha=\sqrt{\frac{\ln N}{7NT}}$. Now let us check the condition for $\beta\leq\frac{1}{2}$, which is equal to $T>\frac{4}{\sqrt{7}}N\ln N$ and holds for $T\to\infty$. Note that $\beta=2\alpha N$ satisfies the condition of $\alpha\frac{N}{\beta}\leq\frac{1}{2}$. We then finish the proof.