Online Conformal Abstention Under Adversarial Bandit Feedback

Minjae Lee[*]
GSAI
POSTECH
minjae.lee@postech.ac.kr

Yoonjae Jung[*]
GSAI
POSTECH
jyjllll1025@postech.ac.kr

Sangdon Park
GSAI & CSE
POSTECH
sangdon@postech.ac.kr

Abstract

As interactive generative systems are increasingly deployed in real-world applications, their tendency to generate unreliable or false responses raises serious concerns. *Conformal abstention* mitigates this risk by ensuring that the system answers only when confident. However, real-world deployments typically provide only partial user feedback (*e.g.,* thumbs up/down) on the selected response and often operate in non-stationary or adversarial environments, for which effective learning methods are largely missing. To bridge this gap, we propose `ExAUL`, a novel online learning framework for conformal abstention with adversarial and partial feedback. Technically, we introduce (i) a novel *conversion lemma* that translates the regret of any bandit algorithm into a selective risk (SR) bound, and (ii) *feedback unlocking*, a strategy that exploits the structure of conformal abstention to extract additional learning signals from partial feedback. We prove that `ExAUL` achieves a regret bound of $\mathcal{O}(\sqrt{T \ln |\mathcal{H}|})$, which translates into an $\mathcal{O}(\sqrt{T})$ bound on SR violation control, matching the controllability of full-information settings despite receiving only partial feedback. While applicable to general generative tasks, we demonstrate the efficacy of `ExAUL` for ensuring the reliability of Large Language Models (LLMs) through empirical validation on question-answering tasks across diverse non-stationary and adversarial settings. Our results demonstrate that `ExAUL` robustly controls the SR while maintaining competitive answering coverage.

1 Introduction

As large language models [37, 32] surpass average human performance, their deployment in real-world human-facing applications has become widespread. However, their tendency to generate unintended or incorrect information raises serious concerns about reliability and safety in high-stakes settings. To address such reliability issues, a common practice is to utilize heuristic uncertainty estimates—such as self-consistency [45, 30] or entropy-based measures [21–23]—and abstain when the uncertainty exceeds a heuristic threshold. While these methods show strong empirical performance, they lack formal guarantees on reliability, where such certification is required for the high-stakes situations in ensuring the trustworthiness of large language models.

A promising approach for providing such guarantees is *selective prediction*, also known as *conformal abstention*, [14, 16, 34, 26, 3, 48], which aims to control an error rate (*e.g.,* the false discovery rate) in a certified manner by abstaining from uncertain predictions. Ideally, the abstainer ensures that whenever it provides an answer, the answer is mostly correct under a specified risk level. However, existing methods [14, 16, 34, 26, 3, 48] that provide such theoretical guarantees are developed under limited *stochastic assumptions*, *i.e.,* data are independently drawn from a fixed distribution. This

[*]Equal contribution

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

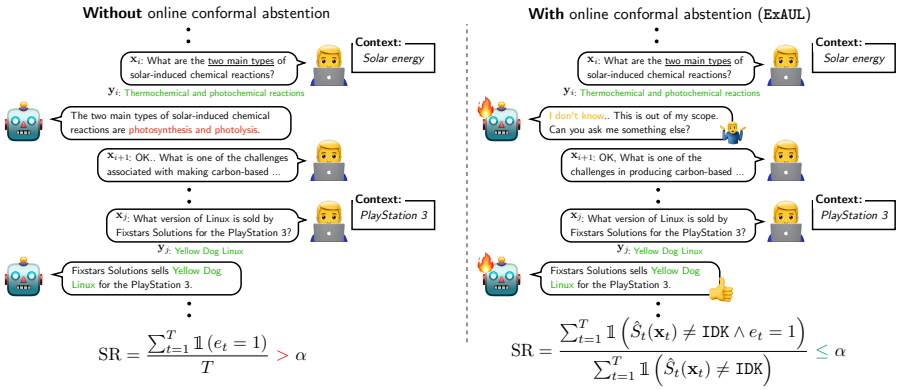


Figure 1: Qualitative examples from an interactive dialog simulation. This demonstrates that our proposed method ExAUL effectively controls the rate of hallucination in the SR by abstaining from answering under a practical online setup with partial feedback. See Section 4.3 for details.

assumption undermines their applicability to dynamic real-world environments involving adversarial distribution shifts. Moreover, these methods require *full feedback* (*i.e.,* ground truth answers), whereas real-world systems typically receive only *partial* or *bandit feedback* (*e.g.,* thumbs-up/down to a generated answer), which is far more practical to obtain. Motivated by these challenges, we ask the following question: "*how can we adaptively learn a certified abstainer with adversarial bandit feedback?*"

To mitigate these limitations, we introduce an *online learning* setup for conformal abstention under *partial* and potentially *adversarial* feedback as learning supervision. We address this problem by proposing a novel online conformal abstention algorithm to control a selective risk (SR), while maximizing selection efficiency (*i.e.,* the ratio of non-abstaining cases). In particular, we first reduce online conformal abstention to bandits for leveraging any regret minimization algorithms, *e.g.,* Exp3-IX [36], for leaning, and then we introduce a *Regret-to-SR conversion lemma* to interpret any regret bounds into SR bounds for guarantees on conformal abstention. Second, we design an algorithm tailored for conformal abstention which fully exploits feedback information. Although partial feedback is practical, it provides an insufficient amount of information compared to full feedback, often requiring more samples for SR guarantees. To address this challenge, we exploit the unique structure of conformal abstention to *unlock* additional information from partial feedback, a technique we call *partial feedback unlocking*. Leveraging this idea, we extend the Exp3-IX for adversarial bandits to online conformal Abstention with adversarially partial feedback UnLocking (ExAUL). Moreover, we provide the $\mathcal{O}(\sqrt{T \ln |\mathcal{H}|})$ expected regret bound of ExAUL —which enables tighter control of the SR guarantee over time—better than a learner with partial feedback, *i.e.,* $\mathcal{O}(\sqrt{T|\mathcal{H}| \ln |\mathcal{H}|})$, and compatible to that with full feedback, *i.e.,* $\mathcal{O}(\sqrt{T \ln |\mathcal{H}|})$, where T is a time horizon and $\mathcal{H}$ is a set of hypotheses of abstainers. This provides our SR guarantee by a desired level α at the rate of $\mathcal{O}(\sqrt{\ln |\mathcal{H}|/T})$ due to our conversion lemma. Finally, we empirically evaluate the efficacy of ExAUL. We consider two tasks (*i.e.,* QA over TriviaQA [20] and Natural Question [25]; and dialog conversations via two dialog agents), four learning environments (*i.e.,* stochastic, distribution-shifted, interactive, and adversarial), and two language models (*i.e.,* GPT-5.4 [37] and LLaMA3.1 [32]), demonstrating that our method (1) controls a desired SR and (2) maintains reasonably low selection inefficiency. See Figure 1 for qualitative results controlling a SR in the interactive environment.

1.1 Related Work

Here, we introduce closely related literature, including selective prediction, conformal abstention, online learning, and bandit problems. See Appendix B for additional related work.

Heuristic-Based Abstention. Recent approaches utilize uncertainty metrics, such as self-consistency [45, 30] and likelihood-/entropy-based scores [21–23], to detect hallucinations. While these metrics serve as effective signals for abstention, the decision rule typically relies on empirical thresholds lacking theoretical justification, thereby failing to provide formal guarantees on error rates (*e.g.,* SR).

Selective Prediction and Conformal Abstention. In high-stakes domains such as medical diagnosis, rejecting an answer rather than making a risky prediction is critical for system reliability [11]. Selective prediction addresses this by abstaining from uncertain predictions to control the error rate in a certified manner. Geifman and El-Yaniv [14] propose selective classification, which learns a threshold for a scoring function to abstain from prediction and controls an SR. Recent advances in conformal prediction provide interpretation of selective prediction in the lens of conformal prediction. Angelopoulos et al. [3] proposes a general risk control framework for conformal prediction, including the SR of a selective classifier as the risk definition. Yadkori et al. [48] consider an un-normalized surrogate of the SR as the risk definition, where it coined the term of *conformal abstention*. Jin and Candès [19] develops a rigorous method to control the SR of a given selected subset from a large candidate pool (also called screening). Extensions of these ideas include exploiting hierarchical labels in selective classification [16] or adapting to generation tasks [34, 48, 26]; Lee et al. [26] generalizes selective prediction to open-ended generation by introducing entailment sets with SR guarantees. However, these prior works predominantly focus on *batch learning* (requiring a calibration set) under *stochastic assumptions* (*i.e.,* i.i.d. data), which limits their applicability in real-world scenarios with distribution shifts. In contrast, we consider a non-stochastic setup that allows for arbitrary distribution shifts and adversarial feedback, proposing a robust online conformal abstention method.

Online Learning and Bandit Problems. Sequential prediction [7, 35, 40, 12] designs learners that adapt to sequentially arriving data. Unlike stochastic learning, online learning methods, *e.g.,* exponential weighting [27], make minimal distributional assumptions and analyze worst-case, possibly adaptive sequences, with guarantees in terms of regret. In a full feedback online learning setting, the learner observes the loss of every prediction each round. In contrast, bandit problems, such as multi-armed bandits [1] and adversarial bandits [4, 6], consider a partial-feedback setting, where the learner only observes feedback of a chosen arm each round. A well-known algorithm is `Exp3` [4], which leverages exponential weighting [27] with an importance-weighted unbiased loss estimator. However, `Exp3` suffers from a dishearteningly large variance of the regret, leading to looser concentration. To obtain high-probability bounds, variants introduce exploration explicitly or implicitly, *i.e.,* `Exp3.P` [4, 6] explicitly mixes the uniform distribution and `Exp3-IX` [36] implements implicit exploration by adding an exploration parameter to the loss estimator. To efficiently leverage partial feedback, structured bandits [39] exploit the functional structure of arms-to-loss functions and semi-bandits [40] allows to choose a set of arms with fixed size for better learning efficiency. Here, we exploit adversarial bandits to mitigate stochastic assumptions in traditional conformal abstention and to consider learning under partial feedback. Moreover, we leverage a unique structure between arms and loss in conformal abstention to achieve better learning efficiency.

2 Online Risk-Controlled Abstention Under Adversarial Bandit Feedback

We consider online learning of an abstainer under adversarial bandit feedback for language models (see Appendix A for its preliminary). Let $\mathcal{W}$ be a set of tokens, $\mathcal{X} = \mathcal{Y} \coloneqq \cup_{i=0}^{\infty} \mathcal{W}^i$ be a set of (token) sequences, and t be a round index. Then, let $\mathcal{H}$ be a finite class of *abstainers* $\hat{S}_t : \mathcal{X} \to \mathcal{Y} \cup \{\texttt{IDK}\}$. Here, given a generator $G : \mathcal{X} \to \mathcal{Y}$, each abstainer $\hat{S}_t \in \mathcal{H}$ is determined by a selection function $\hat{s}_t : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ and abstains from answering when a generated answer $G(\mathbf{x}_t)$ of a question $\mathbf{x}_t$ is uncertain, *i.e.,* $\hat{S}_t(\mathbf{x}_t) \coloneqq \begin{cases} G(\mathbf{x}_t) & \text{if } \hat{s}_t(\mathbf{x}_t, G(\mathbf{x}_t)) = 1 \\ \texttt{IDK} & \text{otherwise} \end{cases}$, where `IDK` represents "I don't know". At each round t, the learner selects an abstainer $\hat{S}_t \in \mathcal{H}$.

To formalize supervision signals in online learning, we define a feedback function $\mathcal{E}_t : \mathcal{Y} \cup \{\texttt{IDK}\} \to [0, 1]$, which is induced from $(\mathbf{x}_t, \mathbf{y}_t)$ chosen by an adversary at round t, and outputs feedback for an abstainer $\hat{S}_t$, *e.g.,* $\mathcal{E}_t(\hat{S}_t(\mathbf{x}_t)) \coloneqq \mathbb{1}(\hat{S}_t(\mathbf{x}_t) \neq_E \mathbf{y}_t)$. We assume that a learner does not observe a ground-truth $\mathbf{y}_t$ but instead receives *partial feedback* $\mathcal{E}_t(\hat{S}_t(\mathbf{x}_t))$, which we simply denote as e_t. Here, $A \neq_E B$ means that A and B are different in terms of a given correctness relation E, *e.g.,* textual-entailment [5]; thus, $e_t = 0$ if the generation is correct and $e_t = 1$ if it is incorrect or $\hat{S}_t(\mathbf{x}_t) = \texttt{IDK}$ by definition. This setup includes real-world user feedback as a special case, where a user typically provides $e_t = 0$ or 1, *e.g.,* thumbs-up or down, instead of $\mathbf{y}_t$. Note that we consider binary feedback for real world user feedback but e_t can be a real value in $[0, 1]$.

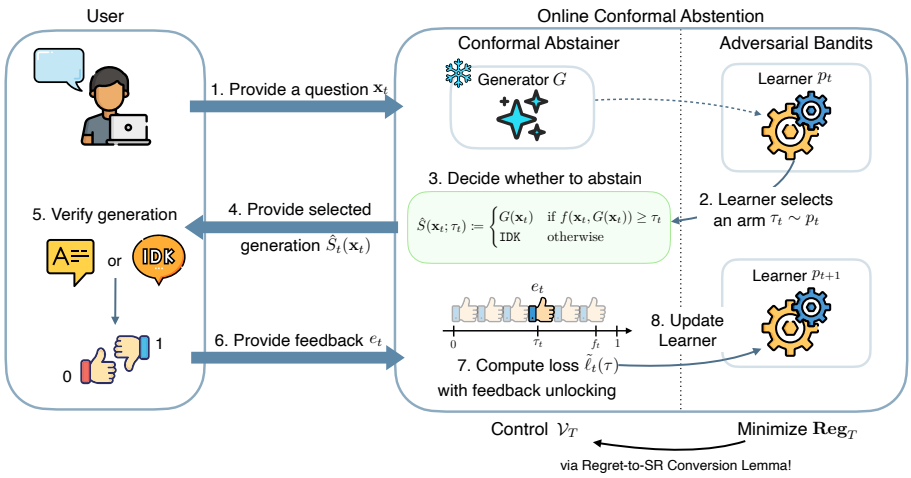


Figure 2: An example of our proposed framework for online conformal abstention. At each round t, (1) the user provides an input $\mathbf{x}_t$, (2) the learner selects an arm $\tau_t \sim p_t$, (3-4) generates $\hat{S}_t(\mathbf{x}_t; \tau_t)$, (5-6) the user provides partial feedback e_t, (7) the loss $\ell_t(\tau_t)$ is computed from e_t, and (8) the learner update p_t by the loss estimator $\tilde{\ell}_t$. Note that the user can be modeled as an adversary.

To learn an online abstainer with this partial feedback, we consider the following online learning protocol with a *non-anticipating adaptive adversary*: each round t until a time horizon T, (1) an adversary chooses $\mathbf{x}_t \in \mathcal{X}$ and $\mathbf{y}_t \in \mathcal{Y}$, which determine the feedback function $\mathcal{E}_t$, (2) a learner observes $\mathbf{x}_t \in \mathcal{X}$ and predicts $\hat{S}_t(\mathbf{x}_t)$, where $\hat{S}_t$ is drawn from a learned distribution p_t over abstainers, and (3) the leaner observes partial feedback e_t and update p_t by using it. Here, we consider an *adaptive adversary*, *i.e.,* $(\mathbf{x}_t, \mathbf{y}_t)$ is drawn from an arbitrary distribution that depends on the interaction history with the learner, *i.e.,* $(\mathbf{x}_1, \mathbf{y}_1), \dots (\mathbf{x}_t, \mathbf{y}_t)$ and $\hat{S}_1(\mathbf{x}_1), \dots, \hat{S}_{t-1}(\mathbf{x}_{t-1})$ but crucially not on the current prediction $\hat{S}_t(\mathbf{x}_t)$, hence *non-anticipating*.

Goal. Under this learning setup, our primary goal is to control risk, where we specifically consider the *selective* risk (SR), of an abstainer at a desired level $\alpha \in [0, 1]$ up to a time horizon T. To this end, we first define the *selective risk violation* as follows:

$$\mathcal{V}_T := \sum_{t=1}^{T} \Big[\mathbb{1}(\hat{S}_t(\mathbf{x}_t) \neq \texttt{IDK}) \cdot e_t - \alpha \mathbb{1}(\hat{S}_t(\mathbf{x}_t) \neq \texttt{IDK}) \Big]. \tag{1}$$

Ideally, we would ensure $\mathcal{V}_T \leq 0$, equivalent to controlling the following SR by α:

$$\mathbf{SR}_T := \frac{\sum_{t=1}^{T} \mathbb{1}(\hat{S}_t(\mathbf{x}_t) \neq \texttt{IDK}) \cdot e_t}{\sum_{t=1}^{T} \mathbb{1}(\hat{S}_t(\mathbf{x}_t) \neq \texttt{IDK})} \leq \alpha, \tag{2}$$

where $\mathbf{SR}_T = \alpha$ if $\hat{S}_t(\mathbf{x}_t) = \texttt{IDK}$ for all t by convention. In the binary-feedback case where $e_t \in \{0, 1\}$, $\mathbf{SR}_T$ is the fraction of incorrect generations among all non-IDK generations, *i.e.,* the false discovery rate (FDR).

As strict control of $\mathcal{V}_T$, *i.e.,* $\mathcal{V}_T \leq 0$, cannot generally be guaranteed, the goal is to learn a distribution p_t over abstainers such that a drawn abstainer $\hat{S}_t \sim p_t$ controls $\mathcal{V}_T$ to be sub-linear in T, *i.e.,* $\mathcal{V}_T \leq \varepsilon(T)$, where $\varepsilon(T) = o(T)$. Accordingly, the learner controls $\mathcal{V}_T$ while the adaptive adversary tries to maximize it. Additionally, SR controllability is the primary objective, but we also wish to minimize selection inefficiency $\mathbf{Ineff}_T := \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(\hat{S}_t(\mathbf{x}_t) = \texttt{IDK})$.

3 ExAUL: Online Conformal Abstention with Feedback Unlocking

Following the conventional conformal and selective prediction literature [44, 14], we instantiate the abstainer class $\mathcal{H}$ of Section 2 through the model-free design of *conformal* abstainers: a scalar

parameterization of $\hat{s}_t$ via a *conformity score* function $f_t : \mathcal{X} \times \mathcal{Y} \to [0, 1)$ measuring the confidence of $G(\mathbf{x}_t)$ being an answer of $\mathbf{x}_t$, *i.e.,* $\hat{s}_t(\mathbf{x}_t, G(\mathbf{x}_t)) \coloneqq \mathbb{1}\left(f_t(\mathbf{x}_t, G(\mathbf{x}_t)) \geq \tau_t\right)$, so the conformal abstainer is parameterized by a threshold τ_t, of often denoted $\hat{S}_t(\cdot; \tau_t)$. We take τ_t from the finely-quantized grid $\mathcal{H} = \{k/(H-1) : k = 0, \ldots, H-1\}$, where H is the number of hypotheses, and identify each $\hat{S}_t$ with its threshold τ_t.

We leverage a regret perspective from bandit problems to design a learning algorithm for conformal abstention that controls the SR under adversarial bandit feedback. After recalling regret minimization in adversarial bandits (Section 3.1), we reduce online conformal abstention to an adversarial bandit (Section 3.2), which enables us to turn any regret bound into an SR bound through a novel conversion lemma (Section 3.3). However, simply leveraging existing bandit algorithms is sample-inefficient under partial feedback. We therefore extend `Exp3-IX` [36] by exploiting the structure of conformal abstainers—*feedback unlocking*—for a tighter regret, and hence a tighter SR, bound (Section 3.4). Adversarial *full* feedback is handled analogously via exponential weighting (`EW`, Appendix C.1).

3.1 Regret Minimization and Adversarial Bandits

The adversarial bandit problem [4] is formulated as an interactive game between a learner and an adversary. At each round $t \in 1, \ldots, T$, the *learner* selects an arm $\tau_t \in \mathcal{H}$ among a set of arms $\mathcal{H}$, while the *adversary* simultaneously chooses a loss function $\ell_t : \mathcal{H} \to \mathbb{R}_{\geq 0}$. The learner then observes only the incurred loss $\ell_t(\tau_t)$ as partial feedback. The learner's performance is evaluated by *regret* [6]

$$\mathbf{Reg}_T \coloneqq \sum_{t=1}^{T} \ell_t(\tau_t) - \min_{\tau \in \mathcal{H}} \sum_{t=1}^{T} \ell_t(\tau), \tag{3}$$

and the goal of the learner is to minimize $\mathbf{Reg}_T$ so that it grows sublinearly in T. Here, the *adaptive adversary* maximizes $\mathbf{Reg}_T$ by choosing ℓ_t based on the previously chosen arms by the learner, *i.e.,* $\tau_1, \ldots, \tau_{t-1}$. This makes $\mathbf{Reg}_T$ a random variable, so we analyze high-probability or expected regret bounds. For further details on regret minimization and an adversarial bandit method (*e.g.,* `Exp3-IX` [36]), see Appendix A.3 and A.4.

3.2 Reduction: From Conformal Abstention to Adversarial Bandits

Here, we map the components of online conformal abstention with partial feedback to those of adversarial bandits to leverage existing algorithms and regret bounds. In particular, a parameter $\tau \in \mathcal{H}$ of our learner corresponds to a bandit arm in adversarial bandits.

Table 1: From online conformal abstention to adversarial bandits

	online conformal abstention	adversarial bandits
models	conformal abstainers $\mathcal{H}$	finite arms $\mathcal{H}$
feedback	e_t	$\ell_t(\tau_t, \alpha)$
metric	$\mathcal{V}_T$ and $\mathbf{Ineff}_T$	$\mathbf{Reg}_T$

We first define a special loss $\ell_t(\tau, \alpha)$ and connect this to feedback e_t. In particular, we introduce the following two key components of loss: $a_t(\tau) \coloneqq \mathbb{1}(\hat{S}(\mathbf{x}_t; \tau) = \texttt{IDK}) \in \{0, 1\}$, which measures the selection inefficiency at time t, called *inefficiency loss*, and $d_t(\tau, \alpha) \coloneqq \mathbb{1}(\hat{S}(\mathbf{x}_t; \tau) \neq \texttt{IDK}) \cdot e_t - \alpha \mathbb{1}(\hat{S}(\mathbf{x}_t; \tau) \neq \texttt{IDK}) + \alpha \in [0, 1]$, which measures the SR violation at time t with a margin by α to penalize the `IDK` response, called *SR loss* with a margin. Based on these, we define the loss function for bandits as follows:

$$\ell_t(\tau, \alpha) \coloneqq \frac{a_t(\tau) + \lambda d_t(\tau, \alpha)}{1 + \lambda} \in [0, 1], \tag{4}$$

where $\lambda \in \mathbb{R}_{>0}$ is a hyperparameter that controls the trade-off between the SR loss and the efficiency loss. Note that from the bandits' perspective, the learner receives feedback in two steps: get feedback e_t from the adversary and uses this to compute the loss $\ell_t(\tau_t, \alpha)$ for the chosen arm τ_t.

Finally, the link between two metrics for online conformal abstention and adversarial bandits remains to be established. In the following section, we propose a novel conversion lemma to connect $\mathbf{SR}_T$ and $\mathbf{Ineff}_T$ to $\mathbf{Reg}_T$. See Figure 2 for an illustration of our full learning pipeline from online conformal abstention to adversarial bandits and back again.

Algorithm 1 `Exp3-IX` for Online Conformal Abstention with Feedback Unlocking (`ExAUL`)

procedure EXAUL($T, \mathcal{H}, \alpha, \lambda, \eta, f, G, \gamma$)
  $w_1(\tau) \leftarrow 1/|\mathcal{H}|$ for all $\tau \in \mathcal{H}$
  for $t = 1, \dots, T$ **do**
    Choose $\tau_t \sim p_t(\tau) = \frac{w_t(\tau)}{\sum_{\bar{\tau} \in \mathcal{H}} w_t(\bar{\tau})}$
    Observe $\mathbf{x}_t$
    Observe e_t
    $\tilde{\ell}_t(\cdot, \alpha) \leftarrow$ UNLOCKING($\mathcal{H}, \alpha, \lambda, \gamma, \tau_t, \mathbf{x}_t, e_t$) (▷) (8)
    $w_{t+1}(\tau) \leftarrow w_t(\tau) \exp(-\eta \tilde{\ell}_t(\tau, \alpha))$ for all $\tau \in \mathcal{H}$
procedure UNLOCKING($\mathcal{H}, \alpha, \lambda, \gamma, \tau_t, \mathbf{x}_t, e_t$)
  $f_t \leftarrow f(\mathbf{x}_t, G(\mathbf{x}_t))$
  $\mathcal{H}_t(\tau_t) \leftarrow \begin{cases} \{\tau \in \mathcal{H} \mid \tau \leq f_t\} \text{ if } \hat{S}(\mathbf{x}_t; \tau_t) \neq \texttt{IDK}, \\ \{\tau \in \mathcal{H} \mid \tau > f_t\} \text{ otherwise.} \end{cases}$
  $\ell_t(\tau, \alpha) \leftarrow$ COMPUTELOSS($\hat{S}(\mathbf{x}_t; \tau), e_t, \alpha, \lambda$) for all $\tau \in \mathcal{H}_t(\tau_t)$ (▷) Algorithm 2
  $\ell_t(\cdot, \alpha \mid \mathcal{H}_t(\tau_t)) \leftarrow \mathbf{0}$
  $\mathcal{P}_t \leftarrow \gamma + \sum_{\bar{\tau} \in \mathcal{H}_t(\tau_t)} p_t(\bar{\tau})$ (▷) (9)
  $\ell_t(\tau, \alpha \mid \mathcal{H}_t(\tau_t)) \leftarrow \ell_t(\tau, \alpha)/\mathcal{P}_t$ for all $\tau \in \mathcal{H}_t(\tau_t)$ (▷) due to $\mathbb{1}\{\tau \in \mathcal{H}_t(\tau_t)\} = 1$
  return $\ell_t(\cdot, \alpha \mid \mathcal{H}_t(\tau_t))$

3.3 Regret-to-SR Conversion

We mainly leverage algorithms that minimize $\mathbf{Reg}_T$ for learning conformal abstainers. Yet, the key requirement for conformal abstention is the SR guarantee of a learner at a desired level. To this end, we introduce a novel perspective on the connection of a regret bound to an SR bound. In particular, *any learner*, including learners for bandits that minimizes the regret (3) with the loss (4), controls the SR (1). This is achievable as the designed loss (4) penalizes when *SR loss* at each t is higher than α, weighted by a parameter λ. See Appendix H.3 for a proof and details on choosing λ.

Lemma 1. *Let $T \in \mathbb{N}$ and $\alpha \in (0, 1)$. For any $(\mathbf{x}_t, \mathbf{y}_t)$ sequences, leading to any loss sequences ℓ_t of (4) with any $a_t : \mathcal{H} \to [0, 1]$ and $e_t \in [0, 1]$, we have*

$$\mathcal{V}_T \leq \frac{T(1 - \textbf{\textit{Ineff}}_T) + (1 + \lambda)\textbf{\textit{Reg}}_T}{\lambda} = \mathcal{O}(T/\lambda + \textbf{\textit{Reg}}_T) = \mathcal{O}(\sqrt{T}), \tag{5}$$

where the second equality (5) holds if we take $\lambda = \sqrt{T}$ and $\textbf{\textit{Reg}}_T = \mathcal{O}(\sqrt{T})$.

Intuitively, as $\lambda \to \infty$ the loss (4) reduces to the risk term d_t and (5) becomes to $\mathcal{V}_T \leq \mathbf{Reg}_T$. Thus, controlling regret directly controls the SR violation. Importantly, this lemma is applicable to learning for both full and partial feedback, as it is agnostic to feedback mechanisms. Also, the lemma implies that if $\mathbf{Reg}_T$ has a sublinear bound, satisfied by most bandit learners, then $\frac{1}{T}\mathcal{V}_T \leq \mathcal{O}(1/\sqrt{T})$, under our adaptive adversarial environment, *i.e.,* an adversary maximizes $\mathcal{V}_T$.

The Connection between $\mathcal{V}_T$ and $\mathbf{SR}_T$. Dividing by $T(1 - \mathbf{Ineff}_T)$ if $\mathbf{Ineff}_T \neq 1$, (5) becomes

$$\mathbf{SR}_T - \alpha \leq \frac{1}{\sqrt{T}} + \frac{(1 + \sqrt{T})\mathbf{Reg}_T}{T\sqrt{T}(1 - \mathbf{Ineff}_T)}, \tag{6}$$

which converges to $\mathcal{O}(1/\sqrt{T})$ if $\mathbf{Ineff}_T$ does not converge to 1, where the condition holds in general (see Appendix I for discussion).

3.4 `ExAUL`: `Exp3-IX` for Online Conformal Abstention with Feedback Unlocking

As for the adversarial bandit learner, we leverage `Exp3-IX` [36], which builds on `EW` [27] for adversarial bandit feedback and provides guarantees against *adaptive* adversaries. The key challenge in moving from full to partial feedback is that limited feedback makes sample complexity of learning high. To address this, we exploit the unique structure of the feedback function $\mathcal{E}_t$ for conformal abstainers (4) to unlock feedback of other arms by a chosen arm, dubbed by *feedback unlocking*. This is closely related to bandits with side observations [31, 36, 28], but the observation structure here

is induced by conformal abstainers (Appendix B). In particular, by the monotonicity of a selection function in τ, *i.e.,* $\hat{s}(\mathbf{x}) \coloneqq \mathbb{1}(f_t \geq \tau)$ where $f_t \coloneqq f(\mathbf{x}_t, G(\mathbf{x}_t))$, the following equality relation between the observed feedback e_t of a chosen arm τ_t and feedback of other arms τ:

$$e_t \coloneqq \mathcal{E}_t(\tau_t) = \begin{cases} \mathcal{E}_t(\tau) \text{ for } \tau \leq f_t & \text{if } f_t \geq \tau_t \\ \mathcal{E}_t(\tau) \text{ for } \tau > f_t & \text{if } f_t < \tau_t \end{cases}. \tag{7}$$

This means that we can reuse the observed feedback e_t of a chosen arm τ_t for all unchosen arm τ for $\tau \leq f_t$ if the abstainer does not reject an answer, *i.e.,* $f_t \geq \tau_t$.

To exploit this feedback unlocking within `Exp3-IX`, we propose a novel loss estimator, which estimates loss over all arms τ given unlocked feedback $\mathcal{H}_t(\tau_t)$ via importance weighting:

$$\ell_t(\tau, \alpha \mid \mathcal{H}_t(\tau_t)) \coloneqq \frac{\ell_t(\tau, \alpha)}{\gamma_t + \sum_{\bar{\tau} \in \mathcal{H}_t(\tau_t)} \mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau})) \cdot p_t(\bar{\tau})} \cdot \mathbb{1}(\tau \in \mathcal{H}_t(\tau_t)) \in [0, \infty), \tag{8}$$

where $\gamma_t > 0$ and $\mathcal{H}_t(\tau_t)$ is a set of feedback unlocked arms by the τ_t's feedback, *i.e.,* $\mathcal{H}_t(\tau_t) \coloneqq \{\tau \in \mathcal{H} \mid \tau \leq f_t\}$ if $f_t \geq \tau_t$ and $\mathcal{H}_t(\tau_t) \coloneqq \{\tau \in \mathcal{H} \mid \tau > f_t\}$ otherwise; see Figure 38 for the visualization of $\mathcal{H}_t(\tau)$. With learning rate $\eta_t > 0$, Algorithm 1 presents our extension of `Exp3-IX` with partial feedback unlocking. Note that our algorithm can be reduced to `Exp3-IX` by letting $\mathcal{H}_t(\tau) = \{\tau_t\}$ for any τ, recovering the `Exp3-IX` 's loss estimator in (12). Also, when computing the estimator (8) in Algorithm 1, the indicator $\mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau}))$ need not be evaluated since it equals 1 for all $\tau, \bar{\tau} \in \mathcal{H}_t(\tau_t)$ by (47) and (49). The normalizer is then the τ-independent, *i.e.,*

$$\gamma + \sum_{\bar{\tau} \in \mathcal{H}_t(\tau_t)} p_t(\bar{\tau}), \tag{9}$$

which can be precomputed each round from the drawn arm τ_t.

Regret Bound. Improving the proof of `Exp3-IX` along with our novel loss estimator (8), our Algorithm 1 achieves a sublinear regret bound, *i.e.,* $\mathbf{Reg}_T \leq \mathcal{O}(\sqrt{T \ln |\mathcal{H}|})$. See Theorem 4 for the regret bound and Appendix H.4 for a proof on our regret bound.

Interestingly, despite partial feedback, our algorithm achieves the same upper bound as `EW` with full feedback, *i.e.,* $\mathcal{O}(\sqrt{T \ln |\mathcal{H}|})$, due to our loss estimator with feedback unlocking (8) which exploits rich information, whereas `Exp3-IX` suffers from an additional factor of $\sqrt{|\mathcal{H}|}$, as in (13), due to its loss estimator with limited information.

SR Violation Bound. The tighter regret bound of Algorithm 1 further enables tighter control of the SR guarantee over time by Lemma 1. In particular, by plugging our regret bound of Algorithm 1 by Theorem 4 into our conversion lemma in Lemma 1, we have the following SR violation bound.

Theorem 1. *Let $\ell_t(\cdot) \in [0, 1]$ of the form (4). For any $T \in \mathbb{N}$, finite hypotheses $\mathcal{H}$, and $\alpha, \delta \in (0, 1)$, Algorithm 1 with $\eta = 2\gamma = \sqrt{\ln |\mathcal{H}|/T}$ provides the following SR violation bound with probability at least $1 - \delta$:*

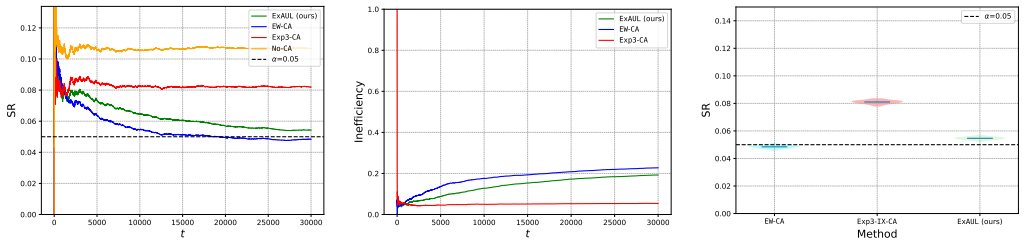
$$\frac{1}{T}\mathcal{V}_T \leq \frac{1 - \textbf{\textit{Ineff}}_T}{\sqrt{T}} + \left(1 + \frac{1}{\sqrt{T}}\right)\left(4\sqrt{\frac{\ln |\mathcal{H}|}{T}} + \left(\frac{1}{T} + \sqrt{\frac{1}{T \ln |\mathcal{H}|}}\right) \ln \frac{2}{\delta}\right). \tag{10}$$

Here, the time-normalized SR violation bound is $\mathcal{O}(\sqrt{\ln |\mathcal{H}|/T})$, which implies that it converges to zero at the rate of $\mathcal{O}(1/\sqrt{T})$ so $\mathbf{SR}_T$ also converges to α to the same rate with a different constant factor, *i.e.,* $\mathbf{SR}_T \leq \alpha + \mathcal{O}(\frac{\sqrt{\ln |\mathcal{H}|/T}}{(1 - \mathbf{Ineff}_T)})$ if $\mathbf{Ineff}_T \neq 1$, as desired.

Remark. *(δ-free parameters). As* (10) *holds for any confidence level $\delta \in (0, 1)$ under a single choice of η and γ independent of δ, Algorithm 1 does not require δ-dependent tuning. Although the $\ln(1/\delta)$ dependence can be improved to $\sqrt{\ln(1/\delta)}$ by tuning η and γ to a fixed confidence level δ, the algorithm is then tied to that δ. We therefore adopt the δ-free choice [36].*

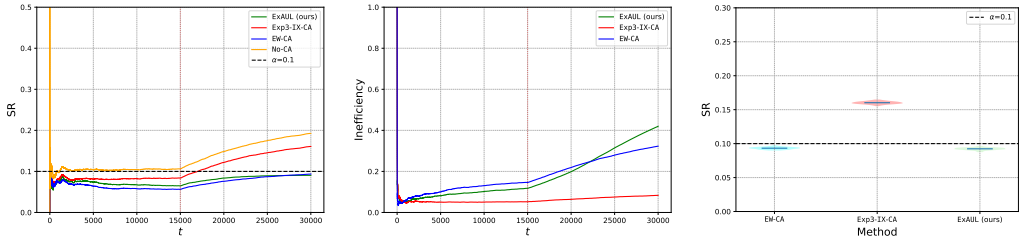
4 Experiments

We empirically justify that `ExAUL` controls the SR while maximizing selection efficiency under four diverse environments:(1) stochastic, (2) distribution-shift, (3) adversarial, and (4) interactive environments.



(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 3: Comparison of conformal abstention methods under a stochastic environment with GPT-5.4 as a generator on TriviaQA ($T = 30\text{K}, \alpha = 0.05$). The violin plots are drawn with randomly chosen 30K samples over 100 random trials.



(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 4: Comparison of conformal abstention methods under a single distribution-shift environment with GPT-5.4 as a generator ($T = 30\text{K}, \alpha = 0.10$), from TriviaQA to NQ. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.

Datasets and Models. We use two datasets for stochastic and distribution-shift environments, 79K Natural Question (NQ) [25] and 93K TriviaQA [20] with two base models, GPT-5.4 [37] and LLaMA3.1-8B-Instruct [32]. To obtain feedback, we employ GPT-5.4-nano to generate feedback conditioned on the $(\mathbf{x}, \mathbf{y}, G(\mathbf{x}))$, which agrees with human annotations on 91.69% of about 4.3K NQ responses from Lee et al. [26]. We simulate distribution-shift environments by mixing two datasets and an adversarial and interactive environment by generating interaction between GPT-5.4 and GPT-5.4-mini models (see Section E.2 and E.3, respectively for details).

Scoring functions. We fix a scoring function as $f_t = f$, considering two $f_{\texttt{std}}$ and $f_{\texttt{con}}$, introduced in Section A.1. In short, $f_{\texttt{std}}$ is a standard log-likelihood score defined as the conditional probability of an answer given a question. $f_{\texttt{con}}$ is a self-consistency score of an answer computed via entailment scores (from an entailment model) across samples. In particular, we compute self-consistency scores over 5 sampled responses [30]. We use $f_{\texttt{con}}$ unless specified.

Methods. We consider two baselines, `Exp3-IX-CA` and `No-CA`, and one performance upper bound algorithm, `EW-CA`. In particular, (1) `Exp3-IX-CA` (Algorithm 6) is an online conformal abstainer with partial feedback by adapting `Exp3-IX` [36] via our regret-to-SR conversion lemma. (2) `No-CA` is a non-conformal abstainer (*i.e.,* $\tau_t = 0$ for all t), to serve as a standard use of a generator without abstention. (3) `EW-CA` (Algorithm 5) is an online conformal abstainer with full feedback. As it exploits full feedback, we use it as our empirical performance upper bound.

Metrics. We measure performance of ours along with comparing methods via the empirical SR, *i.e.,* $\mathbf{SR}_t$ and selection inefficiency, *i.e.,* $\mathbf{Ineff}_t$. Note that `No-CA` achieves zero selection inefficiency by definition, so we do not explicitly add in figures.

Parameters. We set $|\mathcal{H}| = 1\text{K}$ and $\lambda = \sqrt{T}$ by default. Also, a desired SR level α and time horizon T depend on tasks, where α is provided by users.

4.1 Results on Stochastic Environments

We consider a stationary setting where question–answer pairs arrive from a fixed distribution. By default, we follow the common experimental described above. In Figure 3 and 6, we observe

Table 2: SR of ExAUL under imbalanced NQ–TriviaQA streams with NQ fraction p_{NQ} ($T = 60\mathrm{K}$, $\alpha = 0.1$). See Appendix F.3 for details.

		p_{NQ}			
Stream	Model	0.2	0.4	0.6	0.8
Steady	GPT-5.4	0.070	0.087	0.105	0.138
	Llama-3.1-8B	0.082	0.093	0.108	0.130
Discrete	GPT-5.4	0.072	0.083	0.093	0.111
	Llama-3.1-8B	0.084	0.094	0.107	0.124

that ExAUL has clearly better SR controllability to the desired SR compared to Exp3-IX-CA, since Exp3-IX-CA only observes $\ell_t(\tau_t)$, requiring a much longer horizon T (Figure 28). These results demonstrate that our partial feedback unlocking is beneficial in SR control to the desired SR, supporting the regret bound analysis discussed in Section 3.4.

In addition, Figure 26 shows that our algorithm effectively controls the SR across different α. Moreover, Figure 30 and 31 indicate that the inefficiency loss a_t and λ enables the algorithm to control efficient SR guarantees. See additional results in Section F.1.

4.2 Results on Distribution-shift Environments

To evaluate the robustness of the methods, we consider three distribution-shift scenarios: (1) a single instantaneous change by concatenating two datasets, (2) frequent shifts by interleaving fixed-size chunks randomly drawn from each dataset, and (3) progressive shifts by sampling each example according to a linearly increasing mixing probability. See Appendix E.2 for details on shift setups.

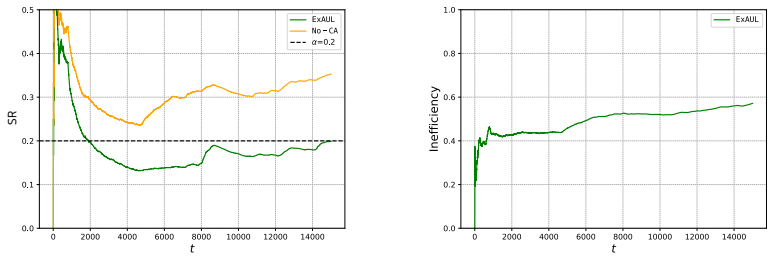
In the TriviaQA-to-NQ shift, Figure 4, 15, and 19 show that Exp3-IX-CA exhibits a sharp increase in SR immediately after the shift, whereas ExAUL maintains SR control. In addition, Figure 27 further shows that ExAUL consistently controls the SR across different target levels α. Moreover, Figure 32 indicates that the regret increases in response to distribution shifts, mirroring the trend of the SR. Additional results across different generators and shift types are provided in Section F.2.

Imbalanced streams. We draw each round from NQ with probability p_{NQ} (steady) or place a TriviaQA block before an NQ block that makes up a fraction p_{NQ} of the stream (discrete). In Table 2, the SR stays below α when TriviaQA dominates and mostly exceeds α once NQ, the harder dataset, forms the majority. The per-dataset SR remains roughly constant across p_{NQ} (about 0.16–0.21 on NQ and 0.05–0.07 on TriviaQA), so the minority dataset shows no additional degradation, although our guarantee concerns the SR over the whole stream rather than per dataset.

Domain and score shifts. Shifting between open-domain QA and the medical benchmark MedQA-USMLE [18] keeps the SR of ExAUL close to α, while moving to MedQA increases inefficiency (Appendix F.2.4). Our analysis also allows f_t to vary over time: when we switch the score from $f_{\mathtt{con}}$ to $f_{\mathtt{std}}$ in the middle of the stream, or replace it with $\mathcal{U}(0,1)$ noise with probability 0.5, ExAUL still brings the SR toward α, at the cost of more abstention and a longer horizon (Appendix F.8).

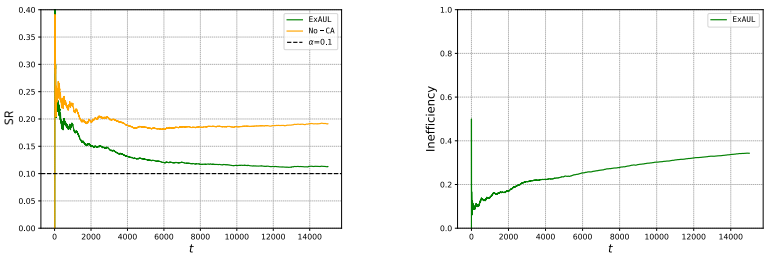
4.3 Results on Agent-Simulated Environments

Adaptive-adversarial environment. In the adaptive setup, a user-acting agent observes the learner's decision history and adaptively selects $(\mathbf{x}_t, \mathbf{y}_t)$ to challenge the learner, consistent with the adaptive-adversary model in Section 3.3. Please refer to Appendix E.4 for the full experimental details and prompts. As illustrated in Figure 5(a) and 5(b), ExAUL demonstrates robust SR control even against a strategy-aware adversary, effectively adapting to shifting adversarial tactics. At the very first stage, the adversary attempts to maximize the risk by exclusively presenting difficult questions. However, this aggressive strategy drives the learner toward a conservative, frequently-abstaining policy, which reduces the SR since abstention incurs zero violation. Consequently, in the later stages, the adversary adapts by mixing incorrect (hard but overconfident) and correct (easy but under-confident) questions to exploit the learner. However, despite this strategic shift creating a highly non-stationary environment, ExAUL continuously adapts its policy, successfully controlling the SR under the target α. This observation empirically validates our theoretical robustness (Lemma 1),



(a) SR (adaptive adversary)

(b) Inefficiency (adaptive adversary)



(c) SR (interactive)

(d) Inefficiency (interactive)

Figure 5: $\mathbf{SR}_t$ and $\mathbf{Ineff}_t$ for ExAUL with GPT-5.4 as an answering agent under (a)-(b) an adaptive-adversarial environment on TriviaQA and NQ ($T = 15\text{K}$, $\alpha = 0.2$) and (c)-(d) an interactive environment ($T = 15\text{K}$, $\alpha = 0.1$).

confirming that ExAUL remains effective even when the adversary actively manipulates the feedback distribution to induce failures.

Interactive environment. We simulate a real-world interactive application, whose question stream is non-stationary, with three agents (Appendix E.3). A user-acting agent (GPT-5.4-mini) asks questions about a context passage from SQuAD [38] and switches to a new passage when it is done or after at most 10 turns, GPT-5.4 answers each question, and an evaluating agent (GPT-5.4-mini) provides the feedback; ExAUL decides whether to abstain from each answer (Figure 1). As shown in Figure 5(c) and 5(d), the SR of ExAUL decreases steadily toward α and reaches about 0.11 at $T = 15\text{K}$, while No-CA remains around 0.19.

5 Conclusion

We propose a learning algorithm for online conformal abstention with adversarial bandit feedback, which provides an SR controllability guarantee by leveraging the adversarial bandit problem. In particular, we reduce conformal abstention to adversarial bandits for leveraging any bandit methods for learning with adversarial bandit feedback and their regret bounds. To interpret the regret bound back to the SR bound, we introduce a novel *Regret-to-SR conversion lemma*, so the any adversarial bandit methods for conformal abstention can be used to control an SR at a desired level. However, the existing bandit methods do not fully exploit the problem structure of conformal abstention so they suffer from sample inefficiency. To address this issue, we propose a novel ExAUL algorithm for adversarial bandits with *partial feedback unlocking*, which addresses the lack of feedback information due to the nature of partial feedback. We theoretically and empirically justify our method by providing an efficient SR violation bound and evaluating diverse learning environments, including stochastic, distribution-shift, interactive, and adaptive adversarial environments.

Limitation. While the SR guarantee holds for any generator and scoring function, its practical efficiency depends on their quality: with a poorly calibrated score or a weak generator, ExAUL may abstain more often and need a longer horizon for the SR to approach α (Appendix I.1). In particular, our method uses the context only through the scoring function f, so extending it to context-aware abstention, *e.g.,* via contextual bandits, is a promising direction to reduce this dependence.

Acknowledgment

We appreciate constructive feedback by anonymous reviewers. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) and the National Research Foundation of Korea (NRF) grant, funded by the Korea government (MSIT) (RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH) (5%); RS-2024-00457882, National AI Research Lab Project (20%); No. RS-2024-00509258 and No. RS-2024-00469482, Global AI Frontier Lab (30%); RS-2025-00560062 (20%); RS-2026-25531812, Safety Shielding(20%); RS-2026-25616313, Star Fellowship (5%)).

References

[1] R. Agrawal. Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in applied probability*, 27(4):1054–1078, 1995.

[2] N. Alon, N. Cesa-Bianchi, C. Gentile, and Y. Mansour. From bandits to experts: A tale of domination and independence. *Advances in Neural Information Processing Systems*, 26, 2013.

[3] A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

[4] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[5] S. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, 2015.

[6] S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[7] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[8] H. Chase. LangChain, Oct. 2022. URL `https://github.com/langchain-ai/langchain`.

[9] H. Chen, Z. Huang, S. Li, and C. Zhang. Understanding bandits with graph feedback. *Advances in Neural Information Processing Systems*, 34:24659–24669, 2021.

[10] T. Chen, Q. Ling, and G. B. Giannakis. An online convex optimization approach to proactive network resource allocation. *IEEE Transactions on Signal Processing*, 65(24):6350–6364, 2017.

[11] C. Cortes, G. DeSalvo, and M. Mohri. Learning with rejection. In *International conference on algorithmic learning theory*, pages 67–82. Springer, 2016.

[12] D. J. Foster and A. Rakhlin. Foundations of reinforcement learning and interactive decision making, 2023. URL `https://arxiv.org/abs/2312.16730`.

[13] H. Ge, H. Bastani, and O. Bastani. Stochastic online conformal prediction with semi-bandit feedback. *arXiv preprint arXiv:2405.13268*, 2024.

[14] Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

[15] I. Gibbs and E. Candès. Adaptive conformal inference under distribution shift, 2021.

[16] S. Goren, I. Galil, and R. El-Yaniv. Hierarchical selective classification. *arXiv preprint arXiv:2405.11533*, 2024.

[17] S. Goren, I. Galil, and R. El-Yaniv. When should llms be less specific? selective abstraction for reliable long-form text generation. *arXiv preprint arXiv:2602.11908*, 2026.

[18] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

[19] Y. Jin and E. J. Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

[20] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147.

[21] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[22] A. Kamath, R. Jia, and P. Liang. Selective question answering under domain shift. *arXiv preprint arXiv:2006.09462*, 2020.

[23] L. Kuhn, Y. Gal, and S. Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

[24] B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International conference on machine learning*, pages 767–776. PMLR, 2015.

[25] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[26] M. Lee, K. Kim, T. Kim, and S. Park. Selective generation for controllable language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=glfYOAzh2f.

[27] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

[28] H. Luo, H. Tong, M. Zhang, and Y. Zhang. Improved high-probability regret for adversarial bandits with time-varying feedback graphs. In *International Conference on Algorithmic Learning Theory*, pages 1074–1100. PMLR, 2023.

[29] M. Mahdavi, R. Jin, and T. Yang. Trading regret for efficiency: online convex optimization with long term constraints. *The Journal of Machine Learning Research*, 13(1):2503–2528, 2012.

[30] P. Manakul, A. Liusie, and M. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[31] S. Mannor and O. Shamir. From bandits to experts: On the value of side-observations. *Advances in neural information processing systems*, 24, 2011.

[32] Meta AI. LLaMA 3.1 8B-Instruct model card. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, 2024.

[33] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 12076–12100, 2023.

[34] C. Mohri, D. Andor, E. Choi, and M. Collins. Learning to reject with a fixed predictor: Application to decontextualization. *arXiv preprint arXiv:2301.09044*, 2023.

[35] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[36] G. Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

[37] OpenAI. GPT-5.4 model card. `https://developers.openai.com/api/docs/models/gpt-5.4`, 2026.

[38] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[39] D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

[40] K. A. Sankararaman. Semi-bandit feedback: A survey of results. *ARXIV CoRR*, 2016.

[41] A. Sid-Ali, I. Lambadaris, Y. Q. Zhao, G. Shaikhet, and A. Asgharnia. Exponentially weighted algorithm for online network resource allocation with long-term constraints. *arXiv preprint arXiv:2405.02373*, 2024.

[42] A. Sinha and R. Vaze. Optimal algorithms for online convex optimization with adversarial constraints. *Advances in Neural Information Processing Systems*, 37:41274–41302, 2024.

[43] I. Stelmakh, Y. Luan, B. Dhingra, and M.-W. Chang. Asqa: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, 2022.

[44] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

[45] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[46] J. Weed, V. Perchet, and P. Rigollet. Online learning in repeated auctions. In *Conference on Learning Theory*, pages 1562–1583. PMLR, 2016.

[47] J. Wei, C. Yang, X. Song, Y. Lu, N. Hu, J. Huang, D. Tran, D. Peng, R. Liu, D. Huang, et al. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827, 2024.

[48] Y. A. Yadkori, I. Kuzborskij, D. Stutz, A. György, A. Fisch, A. Doucet, I. Beloshapka, W.-H. Weng, Y.-Y. Yang, C. Szepesvári, A. T. Cemgil, and N. Tomasev. Mitigating llm hallucinations via conformal abstention, 2024. URL `https://arxiv.org/abs/2405.01563`.

[49] J. Yang, K. Kim, and S. Park. Online conformal prediction with adversarial semi-bandit feedback via regret minimization. *arXiv preprint arXiv:2604.17984*, 2026.

[50] J. Y. Yu and S. Mannor. Unimodal bandits. In *ICML*, pages 41–48, 2011.

[51] Y. Yuan, B. Xu, H. Tan, F. Sun, T. Xiao, W. Li, H. Shen, and X. Cheng. Fact-level calibration and correction for long-form generations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2807–2811, 2025.

[52] C. Zhang, R. Yang, Z. Zhang, X. Huang, S. Yang, D. Yu, and N. Collier. Atomic calibration of llms in long-form generations. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 148–169, 2025.

A Preliminaries

We introduce preliminaries on language generation, selective prediction (or conformal abstention), and adversarial bandits. To this end, let $\mathcal{X}$ be a set of inputs (*e.g.,* examples or questions) and $\mathcal{Y}$ be a set of outputs (*e.g.,* labels or answers).

A.1 Language Generation

We mainly consider language generators as our model to control hallucination. In particular, let $G : \mathcal{X} \to \mathcal{Y}$ be a language generator, where $\mathcal{W}$ is a set of tokens and $\mathcal{X} = \mathcal{Y} \coloneqq \cup_{i=0}^{\infty} \mathcal{W}^i$. Here, each i-th token $\hat{\mathbf{y}}_i$ of a generated answer $\hat{\mathbf{y}} \in \mathcal{Y}$ is decoded from an underlying probability distribution $p(\mathbf{y} \mid \mathbf{x})$, where p is usually learned via language data. For the decoding strategy, we consider greedy decoding, *i.e.,* $\hat{\mathbf{y}}_i = \arg\max_{w \in \mathcal{W}} p(w \mid \mathbf{x}, \hat{\mathbf{y}}_{1:i-1})$, where $\mathbf{y}_{a:b} \coloneqq (\mathbf{y}_a, \dots, \mathbf{y}_b)$. Given a generated answer $\hat{\mathbf{y}}$, there are multiple ways to measure its likelihood of correctness $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. A standard way $f_{\texttt{std}}$ considers a length-normalized token probability, *i.e.,* $f_{\texttt{std}}(\mathbf{x}, \hat{\mathbf{y}}) \coloneqq (p(\hat{\mathbf{y}}_1 \mid \mathbf{x}) \prod_{i=2}^{|\hat{\mathbf{y}}|} p(\hat{\mathbf{y}}_i \mid \mathbf{x}, \hat{\mathbf{y}}_{1:i-1}))^{1/|\hat{\mathbf{y}}|}$. A better alternative is to consider consistency of $\hat{\mathbf{y}}$ to multiple generated answers with sampling [30], *i.e.,* $f_{\texttt{con}}(\mathbf{x}, \hat{\mathbf{y}}) = \hat{\mathbb{E}}_{\mathbf{y}' \sim G(x)} f_E(\mathbf{y}', \hat{\mathbf{y}})$, denoted by a consistency score, where we use sampling for decoding $\mathbf{y}'$, $f_E(\mathbf{y}', \hat{\mathbf{y}})$ measures an entailment score of $\hat{\mathbf{y}}$ given $\mathbf{y}'$ (*i.e.,* whether $\mathbf{y}'$ entails $\hat{\mathbf{y}}$) via an entailment model (*e.g.,* GPT-5.4-nano).

A.2 Selective Prediction and Conformal Abstention

Selective prediction [14] and generation [26] (also known as conformal abstention) provide certified control over the risk of incorrect predictions by abstaining answers (saying `IDK`) when the prediction is uncertain. Given a predictor $\hat{\mathbf{y}} : \mathcal{X} \to \mathcal{Y}$, a selective predictor $\hat{S} : \mathcal{X} \to \mathcal{Y} \cup \{\texttt{IDK}\}$ abstains from returning $\hat{\mathbf{y}}(\mathbf{x})$ if a selection function $\hat{s} : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ deems the prediction uncertain, *i.e.,* $\hat{S}(\mathbf{x}) \coloneqq \begin{cases} \hat{\mathbf{y}}(\mathbf{x}) & \text{if } \hat{s}(\mathbf{x}, \hat{\mathbf{y}}(\mathbf{x})) = 1 \\ \texttt{IDK} & \text{otherwise} \end{cases}$. In learning, the selection function $\hat{s}$ is chosen to possibly satisfy a desired level of a false discovery rate *i.e.,* $\mathbb{P}(\hat{S}(\mathbf{x}) \neq_E \mathbf{y} \mid \hat{S}(\mathbf{x}) \neq \texttt{IDK})$ over the i.i.d. samples of $(\mathbf{x}, \mathbf{y})$. In selective classification [14], the equality relation $\neq_E$ is usually the standard exact matching, which is extended to capture semantic correctness via entailment relation in selective generation [26]. In contrast to the standard stochastic setup as in the previous literature, we consider online learning under partial feedback.

A.3 Regret Minimization

Sequential prediction in multi-armed bandit problems is formulated as an interactive game between a learner and an adversary (also called an environment). In particular, for each time $t \in \{1, \dots, T\}$, a learner selects an arm $\tau_t \in \mathcal{H}$ among a set of arms $\mathcal{H}$ and an adversary selects a loss function $\ell_t : \mathcal{H} \to \mathbb{R}_{\geq 0}$, where a loss associated to an arm τ_t is denoted by $\ell_t(\tau_t)$ and provided to the learner as partial feedback (*i.e.,* only the loss for the selected arm is provided). Note that this interactive game often describes in rewards, but we consider losses, instead of rewards.

The main goal of this interactive game is to find a learner whose performance is as good as the best learner in hindsight. Here, the performance is measured by *regret*, following standard bandits literatures [6], *i.e.,*

$$\mathbf{Reg}_T \coloneqq \sum_{t=1}^{T} \ell_t(\tau_t) - \min_{\tau \in \mathcal{H}} \sum_{t=1}^{T} \ell_t(\tau). \tag{11}$$

Thus, the goal becomes to find a learner that minimizes the regret such that an achieved regret bound is sublinear in T. Here, we say that the adversary is *oblivious* if the sequence of loss functions are chosen regardless of learner's previous selections (or chosen in advance before the game), and the adversary is *adaptive* if the loss function at time t, *i.e.,* ℓ_t, is drawn a distribution that depends on the previously chosen arms, *i.e.,* $\tau_1, \dots, \tau_{t-1}$. In this paper, we consider the adaptive adversary which is stronger than the oblivious one. Moreover, under the adaptive adversary setup, a deterministic learner cannot win the game (*i.e.,* it does not achieve a sub-linear regret bound), so we consider a randomized learner. This implies that τ_t and ℓ_t are two sources of randomness, so $\mathbf{Reg}_T$ is a random variable. Thus, we consider the high-probable regret bound or a bound of an expected regret $\mathbb{E}\mathbf{Reg}_T$. Note that we only consider the learner's randomization in the analysis, since our bounds are derived to hold for any adversary [6, 36].

In the following section, we review one regret minimization algorithm for bandits with an adaptive adversary, called `Exp3-IX`, and see Appendix D for other algorithms.

A.4 Adversarial Bandits and `Exp3-IX` Algorithm

Traditional online learning typically assumes full feedback, *i.e.,* loss for any hypothesis is computable. However, practical applications may only have partial feedback, *i.e.,* loss for a chosen hypothesis is only given. We model this via the adversarial multi-armed bandit problem. The regret is as in online learning, but the learner only receives for a chosen arm, while the sequence $\ell_t(\tau_t)$ is chosen by an oblivious or adaptive adversary that may condition on the history of chosen arms but not on the learner's fresh randomness at time t.

To obtain the high-probability control of the $\mathbf{Reg}_T$ over an adaptive adversary, Exponential-weights for Exploration and Exploitation with Implicit Exploration (`Exp3-IX`)[36] is a natural choice. It employs implicit-exploration (IX) loss estimates within an exponential-weights scheme and does not require explicit mixing with the uniform distribution like `Exp3.P` [4]. In particular, `Exp3-IX` uses the variance-reduced loss estimator for each arm $\tau \in \mathcal{H}$ constructed from the observed loss of the chosen arm τ_t, *i.e.,*

$$\tilde{\ell}_t(\tau \mid \{\tau_t\}) \coloneqq \frac{\ell_t(\tau)}{\gamma_t + p_t(\tau)} \cdot \mathbb{1}(\tau_t = \tau) \in [0, \infty). \tag{12}$$

where $\gamma_t > 0$ is properly selected by an analysis; thus by using (12) in $\mathbf{Reg}_T$, with non-increasing learning rate η_t, we have the following regret with high probability for `Exp3-IX` . See AppendixH.2 for a complete proof.

Theorem 2. *[36] Let $\ell_t(\cdot) \in [0, \ell_{max}]$. For any $T \in \mathbb{N}$, finite arms $\mathcal{H}$, and $\delta \in (0,1)$, Algorithm 4 provides the following regret bound with probability at least$1-\delta$ if $\eta_t = \frac{2\gamma_t}{\ell_{max}} = \sqrt{2\ln|\mathcal{H}|/(\ell_{\max}^2 T|\mathcal{H}|)}$:*

$$\boldsymbol{Reg}_T \leq \ell_{max}\left(2\sqrt{2T|\mathcal{H}|\ln|\mathcal{H}|} + \left(1 + \sqrt{\frac{T|\mathcal{H}|}{2\ln|\mathcal{H}|}}\right)\ln\delta^{-1}\right). \tag{13}$$

We will exploit `Exp3-IX` for learning selective generators with partial feedback under an adaptive adversary. See Appendix C for other online learning and bandit algorithms.

A.5 Online Learning and `EW` Algorithm

Online learning considers to design a learner that adapts to any sequence of full feedback on hypotheses without any or with little assumption on data generation process. In particular, the goal of an online learner is to find a distribution p_t over hypotheses with full feedback that minimizes the regret.

Along with many online learning algorithms, we introduce Exponential Weighting (`EW`) [27], which minimizes the regret with finite hypotheses (Algorithm 3). In particular, `EW` maintains the goodness of weights for each hypothesis in terms of cumulative loss and updates them from feedback. The feedback of a hypothesis τ_t at time t is represented by loss $\ell_t(\tau_t)$ but we can compute this loss for all hypotheses other than a chosen hypothesis τ_t as we directly observe a true label $\mathbb{1}(G(\mathbf{x}_t) =_E \mathbf{y}_t)$ at the same time t. This `EW` algorithm satisfies the following regret bound. See Appendix H.1 for a proof.

Theorem 3. *[27, 12] Let $\ell_t(\cdot) \in [0, \ell_{max}]$. For any $T \in \mathbb{N}$ and finite hypotheses $\mathcal{H}$, Algorithm 3 provides the following expected regret bound if $\eta = \sqrt{8\ln|\mathcal{H}|/(\ell_{max}^2 T)}$:*

$$\sum_{t=1}^{T} \mathbb{E}_{\tau_t \sim p_t} \ell_t(\tau_t) - \min_{\tau \in \mathcal{H}} \sum_{t=1}^{T} \ell_t(\tau) \leq \ell_{max}\sqrt{T\ln|\mathcal{H}|/2} = \mathcal{O}(\ell_{max}\sqrt{T\ln|\mathcal{H}|}), \tag{14}$$

which holds for arbitrary loss sequences, thus, implying that $\mathbb{E}\boldsymbol{Reg}_T$ is also bounded by $\mathcal{O}(\ell_{max}\sqrt{T\ln|\mathcal{H}|})$ as the linearity of expectation and Tower property. Moreover, for any $0 < \delta < 1$, the following regret bound also holds:

$$\sum_{t=1}^{T} \ell_t(\tau_t) - \min_{\tau \in \mathcal{H}} \sum_{t=1}^{T} \ell_t(\tau) \leq \ell_{max}\sqrt{\frac{T\ln|\mathcal{H}|}{2}} + \ell_{max}\sqrt{\frac{T\ln\delta^{-1}}{2}},$$

with probability at least $1-\delta$.

We will leverage and extend this `EW` for learning selective generators with full feedback.

B Additional Related Work

Constrained Online Learning Several works have studied online learning with constraints. Mahdavi et al. [29] first proposes long-term constraints in online convex optimization with time-invariant constraints, which allows to have the constraint function g_t to be violated at each time step, *i.e.,* $g(\mathbf{x}_t) > 0$, but eventually satisfied, *i.e.,* $\sum_{t=1}^{T} g(\mathbf{x}_t) \leq 0$. Some other works solve a similar problem but with time-varying constraints, which is more challenging. Chen et al. [10], Sid-Ali et al. [41] solve a network resource allocation task by viewing this as an constrained optimization problem. Especially, Sid-Ali et al. [41] solves the problem without a convex assumption on loss. Sinha and Vaze [42] gives analysis on optimal algorithms for online convex optimization with adversarial constraints. We similarly consider a SR guarantee to be satisfied over time but with bandit feedback.

Conformal Prediction with Bandit Feedback There is little work about online conformal prediction [15] in bandit settings. Ge et al. [13] proposes an online conformal prediction method with semi-bandit feedback, where we can observe the true label only if it is predicted. In a concurrent work, Yang et al. [49] also explored online conformal prediction under a similar semi-bandit feedback setting. Our method shares the practical motivation of utilizing bandit feedback, but we specifically focus on conformal abstention.

Feedback Unlocking and Structured Bandits. Feedback unlocking is related to structured bandits, which improve sample efficiency by exploiting structure among arms, either in the expected rewards, *e.g.,* unimodality over a graph of arms [50], or in the feedback, *e.g.,* cascading feedback, where a click reveals the attraction of all preceding items [24]. Closest to ours are bandits with feedback graphs [31, 2, 9], where pulling an arm also reveals the losses of its neighbors and the regret scales with graph quantities such as the independence number instead of the number of arms; Neu [36] extends `Exp3-IX` to this setting, and Luo et al. [28] obtain high-probability bounds for time-varying graphs. Inspired by this line of work, we tailor graph feedback to conformal abstention: the graph is not given externally but induced by the monotonic threshold structure, and consists of two cliques, $\{\tau \leq f_t\}$ and $\{\tau > f_t\}$, that change with f_t. Its independence number is thus at most two, which yields regret of the full-information order (Theorem 4), and our estimator (8) takes the form of the IX estimator of Neu [36]. A similar one-sided feedback appears in repeated auctions, where the bidder observes its value only when it wins [46].

C Additional Method

C.1 Online Conformal Abstention with Full Feedback

Here, we specifically leverage the conventional exponential weighting (`EW`) algorithm [27] for online learning under adaptive adversaries to re-purpose it for addressing online conformal abstention with full feedback. This is an oracle method with the performance upper bound of our method with partial feedback. To this end, we provide the reduction of online conformal abstention with full feedback to the online learning problem. Then, we introduce a modified `EW` algorithm for conformal abstention, called `EW-CA`, followed by its regret bound. Finally, we leverage the proposed conversion lemma, which converts the regret bound of any online learning algorithms into the SR bound, to show the SR controllability of `EW-CA` at a desired level.

Reduction: From Online Conformal Abstention to Online Learning. We convert online conformal abstention with full feedback to conventional online learning (See Table 3 for a summary). In conformal abstention, models are defined as a set of conformal abstainers $\hat{S}$ or equivalently a set of selection parameters $\mathcal{H}$, considered as hypotheses $\mathcal{H}$ in online learning. This model is sequentially updated by leveraging online feedback. In particular, we consider feedback whether a generated answer $G(\mathbf{x}_t)$ is semantically incorrect with respect to a true answer $\mathbf{y}_t$, *i.e.,* $\mathbb{1}(G(\mathbf{x}_t) \neq_E \mathbf{y}_t)$. This essentially provides full feedback for any $\tau \in \mathcal{H}$ as it is the same as $\mathbb{1}(\hat{S}(\mathbf{x}_t; \tau) \neq_E \mathbf{y}_t)$ by the definition of $\hat{S}$ if $\hat{S}(\mathbf{x}_t; \tau) \neq \texttt{IDK}$. This feedback is converted into loss in online learning, as defined in (4). Note that d_t in the loss leverages the feedback $\mathbb{1}(G(\mathbf{x}_t) \neq_E \mathbf{y}_t)$ by the definition. Finally, our goal of finding a learning algorithm, which controls the SR by α and minimizes selection inefficiency in online conformal abstention, is reduced to finding an online learning algorithm that minimizes the regret $\mathbf{Reg}_T$ with any loss sequences $\ell_t(\tau, \alpha)$ in (4).

Table 3: From online conformal abstention to online learning

	online conformal abstention	online learning
models	conformal abstainers $\mathcal{H}$	hypotheses $\mathcal{H}$
feedback	$\mathbb{1}(G(\mathbf{x}_t) \neq_E \mathbf{y}_t)$	$\ell_t(\tau, \alpha)$
metric	$\mathcal{V}_T$ and $\mathbf{Ineff}_T$	$\mathbf{Reg}_T$

Algorithm and Its Regret Bound. We re-purpose `EW` for conformal abstention called `EW-CA`. This mainly uses the special loss $\ell_t(\tau, \alpha)$ in (4) for conformal abstention. See Algorithm 5 for details. The corresponding regret bound is directly obtained from Theorem 3 where $\ell_{\max} \coloneqq \max(\lambda, 1 + \lambda\alpha) \leq 1 + \lambda$.

Regret-to-SR Conversion. Once we devise a regret minimization algorithm, *i.e.,* `EW-CA`, the regret guarantee is simply interpreted as the SR guarantee by Lemma 1. This eventually provides the SR bound for `EW-CA`.

D Algorithms

Algorithm 2 Procedure for Computing the Loss (4) with Partial Feedback

procedure ComputeLoss($\hat{S}, e, \alpha, \lambda$)
  $a \leftarrow \mathbb{1}(\hat{S} = \texttt{IDK})$
  $d \leftarrow \mathbb{1}(\hat{S} \neq \texttt{IDK}) \cdot e - \alpha\mathbb{1}(\hat{S} \neq \texttt{IDK}) + \alpha$
  $\ell \leftarrow \frac{a+\lambda d}{1+\lambda}$
  return ℓ

Algorithm 3 Exponential Weighting (`EW`) [27, 12]

procedure EW($T, \mathcal{H}, \eta$)
  $w_1 \leftarrow (1/|\mathcal{H}|, \ldots, 1/|\mathcal{H}|)$
  for $t = 1, \ldots, T$ **do**
    Observe $\mathbf{x}_t$
    Predict $\hat{\mathbf{y}}(\mathbf{x}_t; \tau_t)$ for $\tau_t \sim p_t(\tau) = w_t(\tau) / \sum_{\bar{\tau} \in \mathcal{H}} w_t(\bar{\tau})$
    Observe $\mathbf{y}_t$
    Update $w_{t+1}(\tau) \propto w_t(\tau) \exp(-\eta \ell(\mathbf{y}_t, \hat{\mathbf{y}}(\mathbf{x}_t; \tau)))$ for all $\tau \in \mathcal{H}$

Algorithm 4 Exp3 with Implicit eXploration (`Exp3-IX`) [36]

procedure EXP3-IX($T, \mathcal{H}, \eta_t, \gamma_t$)
  $w_1 \leftarrow (1/|\mathcal{H}|, \ldots, 1/|\mathcal{H}|)$
  for $t = 1, \ldots, T$ **do**
    Choose an arm $\tau_t \sim p_t(\tau) = w_t(\tau) / \sum_{\bar{\tau} \in \mathcal{H}} w_t(\bar{\tau})$
    Observe $\ell_t(\tau_t)$
    $\hat{\ell}_t(\tau) \leftarrow \frac{\ell_t(\tau_t)}{\gamma_t + p_t(\tau)} \mathbb{1}(\tau_t = \tau)$ for all $\tau \in \mathcal{H}$
    Update $w_{t+1}(\tau) \propto \exp(-\eta_t \sum_{s=1}^{t} \hat{\ell}_s(\tau))$ for all $\tau \in \mathcal{H}$

Algorithm 5 Exponential Weighting for Online Conformal Abstention (`EW-CA`)

procedure EW-CA($T, \mathcal{H}, \alpha, \lambda, \eta$)
  $w_1(\tau) \leftarrow 1/|\mathcal{H}|$ for all $\tau \in \mathcal{H}$
  for $t = 1, \ldots, T$ **do**
    Observe $\mathbf{x}_t$
    Predict $\hat{S}_t(\mathbf{x}_t; \tau_t)$ where $\tau_t \sim p_t(\tau) = w_t(\tau) / \sum_{\bar{\tau} \in \mathcal{H}} w_t(\bar{\tau})$
    Observe $\mathbf{y}_t$ thus observe $\mathbb{1}\left(G(\mathbf{x}_t) \neq_E \mathbf{y}_t\right)$
    $a_t(\tau) \leftarrow \mathbb{1}(\hat{S}(\mathbf{x}_t; \tau) = \texttt{IDK})$
    $d_t(\tau, \alpha) \leftarrow \mathbb{1}(\hat{S}(\mathbf{x}_t; \tau) \neq \texttt{IDK} \wedge G(\mathbf{x}_t) \neq_E \mathbf{y}_t) - \alpha \mathbb{1}(\hat{S}(\mathbf{x}_t; \tau) \neq \texttt{IDK}) + \alpha$
    $\ell_t\left(\tau, \alpha\right) \leftarrow \frac{a_t(\tau) + \lambda d_t(\tau, \alpha)}{1 + \lambda}$ for all $\tau \in \mathcal{H}$
    Update $w_{t+1}(\tau) \propto w_t(\tau) \exp\left\{-\eta \cdot \ell_t\left(\tau, \alpha\right)\right\}$ for all $\tau \in \mathcal{H}$

Algorithm 6 `Exp3-IX` for Online Conformal Abstention with Partial Feedback (`Exp3-IX-CA`)

procedure EXP3-IX-CA($T, \mathcal{H}, \alpha, \lambda, \eta, \gamma$)
  $w_1(\tau) \leftarrow 1/|\mathcal{H}|$ for all $\tau \in \mathcal{H}$
  for $t = 1, \ldots, T$ **do**
    Observe $\mathbf{x}_t$
    Predict $\hat{S}_t(\mathbf{x}_t; \tau_t)$ where $\tau_t \sim p_t(\tau) = \frac{w_t(\tau)}{\sum_{\bar{\tau} \in \mathcal{H}} w_t(\bar{\tau})}$
    Observe e_t
    $\ell_t\left(\tau, \alpha\right) \leftarrow \textsc{ComputeLoss}(\hat{S}(\mathbf{x}_t; \tau), e_t, \alpha, \lambda)$ for τ_t
    $\ell_t\left(\tau, \alpha \mid \{\tau_t\}\right) \leftarrow \frac{\ell_t(\tau, \alpha)}{\gamma + p_t(\tau)} \cdot \mathbb{1}(\tau_t = \tau)$ for all $\tau \in \mathcal{H}$
    Update $w_{t+1}(\tau) \propto \exp(-\eta \sum_{s=1}^{t} \ell_s\left(\tau, \alpha \mid \{\tau_t\}\right))$ for all $\tau \in \mathcal{H}$

E Experiment Setup

E.1 Computing Environment

We use 4 NVIDIA A100 80GB with 128 CPUs for experiments.

E.2 Distribution-shift Environment Setup

We describe the experimental setup used to evaluate the robustness of conformal abstention by three distinct distribution-shift scenarios. We detail the chunking strategy and the procedure for constructing each shift environment in the following.

Single Shift. The dataset is composed of two large homogeneous segments, each containing 15K examples randomly sampled from NQ and TriviaQA. A single distribution shift occurs at the boundary between these two segments.

Alternating Shift. The dataset constructed by a sequence of ten 3K-example chunks that alternate between NQ and TriviaQA (e.g., NQ, TriviaQA, NQ, TriviaQA and so forth). Each chunk induces a distribution transition at its endpoint, resulting in frequent and periodic shifts.

Gradual Shift. This setup generates a smooth distribution change over time by drawing individual samples probabilistically between NQ and TriviaQA, with the sampling probability gradually shifting from pure NQ (or TriviaQA) at the start to pure TriviaQA (or NQ) at the end. In particular, given a multinomial distribution $[\frac{t}{T}, 1 - \frac{t}{T}]$ for each dataset that changes over time, we first randomly choose a dataset which follows this multinomial distribution, and then we sample one question-answering pair in the sampled dataset. Over the full sequence length, this creates linear interpolation between the two datasets, enabling analysis of model performance under a progressive distribution shift.

E.3 Interactive Environment Setup

For the interactive environment, we consider dialog-based conversations. In particular, GPT-5.4 serves as a question-answering agent, while we utilize two GPT-5.4-mini APIs, where one is a user-acting agent to generate questions, and another is an evaluating agent, having access to the responses.

To instruct the user-acting agent to generate questions, we use the SQuAD [38] dataset, which consists of many question-answer pairs, each associated with a gold context. Here, we discard the QA pairs and only use $20,962$ context passages, which vary over a range of topics (*e.g.,* science, history, celebrities, entertainment, etc.), forming a diverse set of natural distributions. During the simulation, we randomly select a context from the dataset and provide it to the user-acting agent, which is prompted to continuously generate questions based on the given context. (*e.g.,* Figure 1). To mimic a dynamically shifting environment, *e.g.,* the user's interest shifts over time, we further instruct the user-acting agent to change the context during the interaction when it determines it is done with the current one (*e.g.,* `"When you are done with the current context, say 'SHIFT' to request a new context."`) or after at most 10 question-answering turns. To design this custom workflow, we implement the core functions of the simulation, including memory handling and multi-turn conversation, using LangGraph, a framework built on LangChain [8].

E.4 Adaptive-Adversary Environment Setup

We simulated a strategy-aware adaptive adversary (GPT-5.4-mini) targeting a learner employing GPT-5.4 as the generator, over the combined NQ and TriviaQA datasets. In this setup, a user-acting agent observes the learner's decision history and strategically selects questions to maximize the learner's failure—specifically targeting accepted incorrect answers and rejected correct answers—in alignment with the adaptive adversary defined in Section 3.3.

Note that this simulation employs an LLM assigned an adversarial role to actively induce failures, serving as a realistic proxy rather than a theoretically optimal adversary (which implies omniscience). In other words, this black-box adversary does not possess full parametric knowledge of the learner—such as the specific algorithm (`ExAUL`) or the scoring function f—but relies solely on observed decisions, and the prompt that the learner adapts over time. The specific prompts used to enforce this adversarial objective are provided below.

```
# System Prompt
You are an adversary engaged in a game with a learner model. Your goal is to expose
    the learner's weaknesses.

Below is the list of candidate questions you can ask, with the decision history:
{history}

Current Status:
- the current step: {step}
- The total mean of incorrect answers not rejected (IDK==False and label==0) over
    steps: {total_incorrect_not_rejected}
- The total mean of correct answers rejected (IDK==True and label==1) over steps: {
    total_correct_rejected}

{instructions}

When you want to find more proper questions for maximizing your objective, say '
    SHIFT' to get a new question.
You must decide this **yourself** - do **not** ask whether to 'SHIFT', simply
    respond with 'SHIFT' by yourself when needed.
```

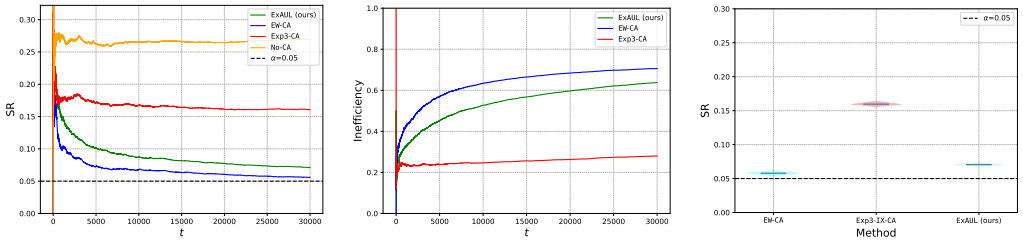
```
# Instructions
**Your Objective:**
You are an adversary trying to **maximize the total mean of incorrect answer not
    rejected (IDK==False and label==0)** by the learner.
However, you must know that the learner trains a threshold to reject incorrect
    answers based on your questions and their confidence score of answers,
thus you need to be strategic in asking questions, i.e., you should also **maximize
    the total mean of correct answers rejected (IDK==True and label==1)** by the
    learner.
E.g., if the learner starts to reject incorrect answers more often, you should also
    try to ask other questions that may fool the learner.

Keep in mind that:
- The learner updates its rejection threshold over time.
- Look for IDK patterns in the history list.
- Select questions that are likely to maximize your objective.

**Output Format:**
- If you select a question, respond with **ONLY the index number** (e.g., 5).
- Do not write any other explanation or text.
```

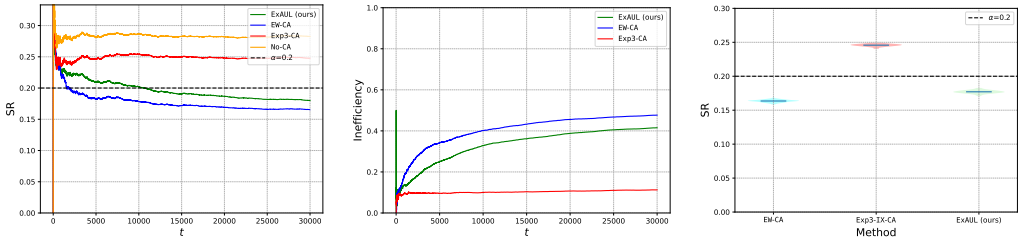
F Additional Experiments

F.1 Stochastic Environment



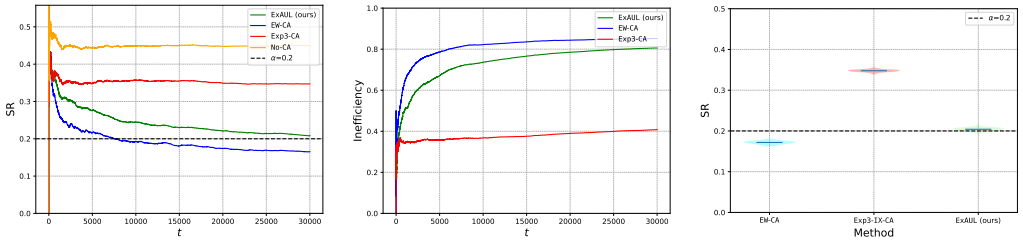
(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 6: Comparison of conformal abstention methods under a stochastic environment with LLaMA3.1-8B-Instruct as a generator on TriviaQA ($T = 30\text{K}, \alpha = 0.05$). The violin plots are drawn with randomly chosen 30K samples over 100 random trials.



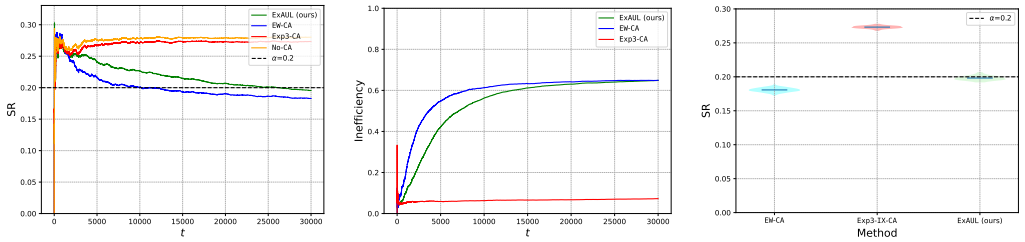
(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 7: Comparison of conformal abstention methods under a stochastic environment with GPT-5.4 as a generator on NQ ($T = 30\text{K}, \alpha = 0.20$). The violin plots are drawn with randomly chosen 30K samples with 100 random trials.



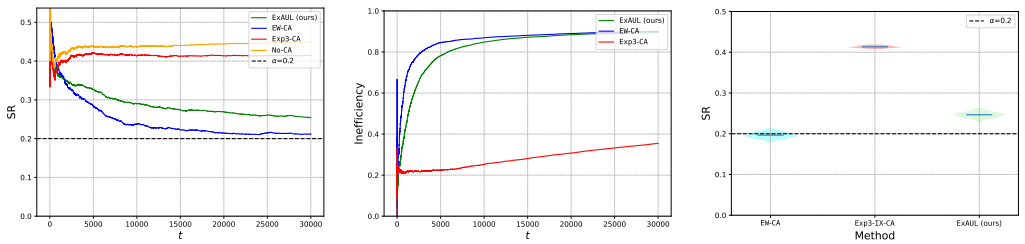
(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 8: Comparison of conformal abstention methods under a stochastic environment with LLaMA3.1-8B-Instruct as a generator on NQ ($T = 30\text{K}, \alpha = 0.20$). The violin plots are drawn with randomly chosen 30K samples with 100 random trials.



(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 9: Comparison of conformal abstention methods under a stochastic environment with GPT-5.4 as a generator on NQ ($T = 30\text{K}, \alpha = 0.20$) and $f_{\texttt{std}}$ as our scoring function. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.

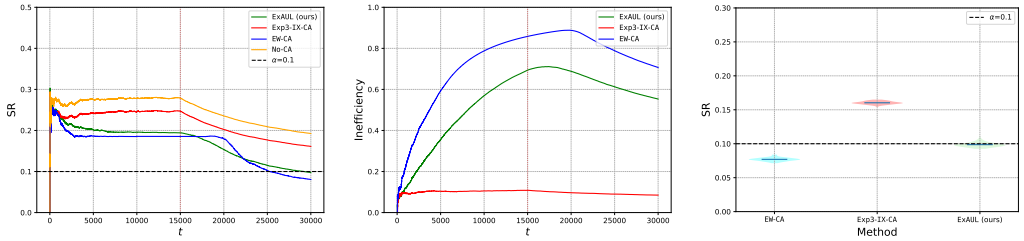


(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 10: Comparison of conformal abstention methods under a stochastic environment with LLaMA3.1-8B-Instruct as a generator on NQ ($T = 30\text{K}, \alpha = 0.20$) and $f_{\texttt{std}}$ as our scoring function. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.

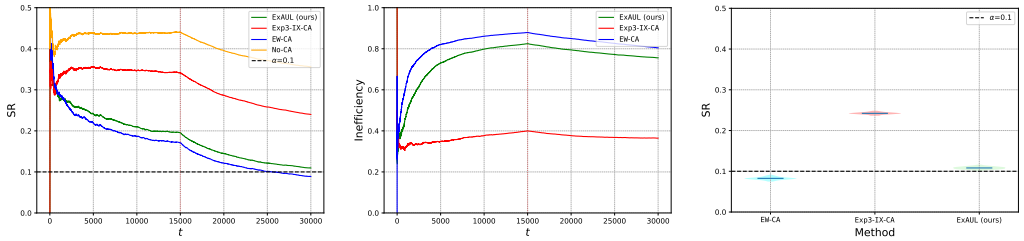
F.2 Distribution-shift Environment

F.2.1 Single Shift



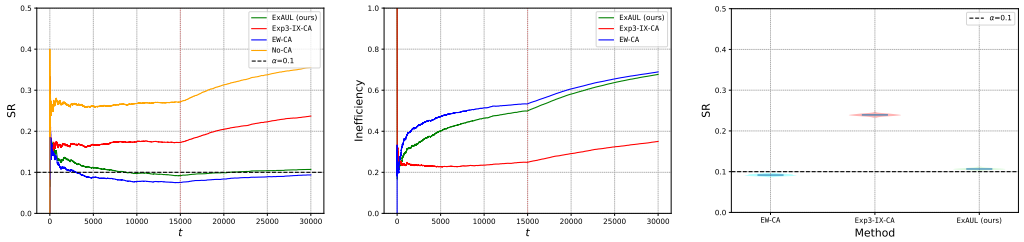
(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 11: Comparison of conformal abstention methods under a distribution-shift environment with GPT-5.4 as a generator ($T = 30\text{K}, \alpha = 0.10$). We consider a single distribution shift from NQ to TriviaQA. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.



(a) SR over time (b) Inefficiency over time (c) SR distribution

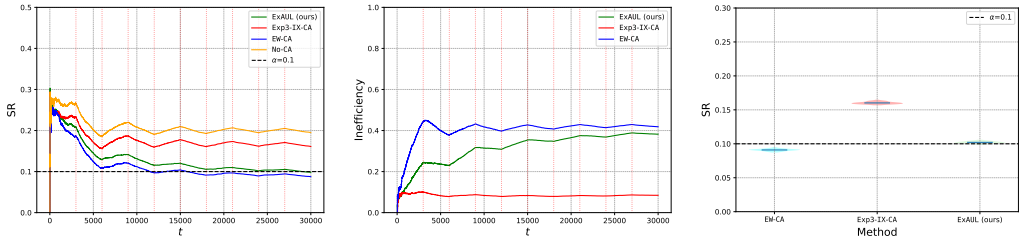
Figure 12: Comparison of conformal abstention methods under a distribution-shift environment with LLaMA3.1-8B-Instruct as a generator ($T = 30\text{K}, \alpha = 0.10$), in a single distribution shift from NQ to TriviaQA. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.



(a) SR over time (b) Inefficiency over time (c) SR distribution

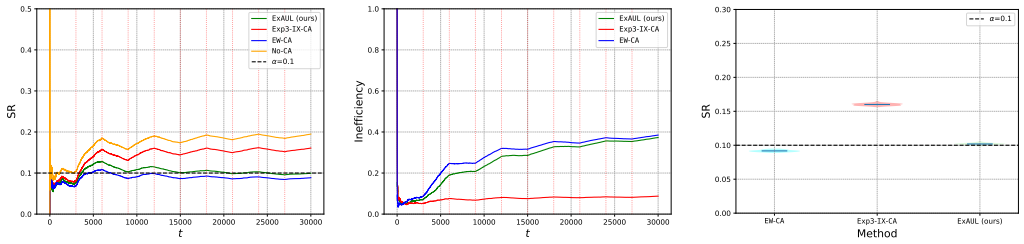
Figure 13: Comparison of conformal abstention methods under a distribution-shift environment with LLaMA3.1-8B-Instruct as a generator ($T = 30\text{K}, \alpha = 0.10$), in a single distribution shift from TriviaQA to NQ. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.

F.2.2 Alternating Shift



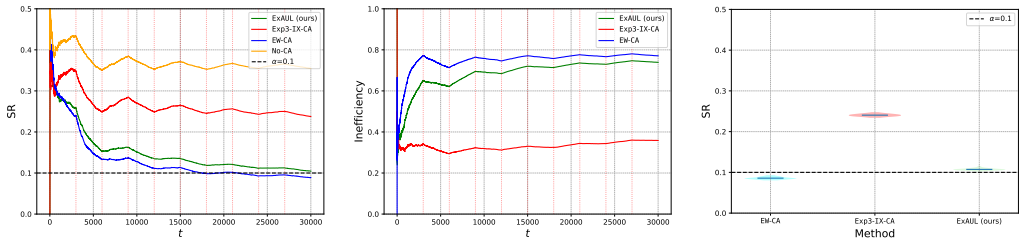
(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 14: Comparison of conformal abstention methods under a distribution-shift environment with GPT-5.4 as a generator ($T = 30\text{K}, \alpha = 0.10$), in a distribution shift by alternating between NQ and TriviaQA over time, beginning with NQ. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.



(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 15: Comparison of conformal abstention methods under an alternating distribution-shift environment with GPT-5.4 as a generator ($T = 30\text{K}, \alpha = 0.10$), alternating between TriviaQA and NQ, starting with TriviaQA. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.



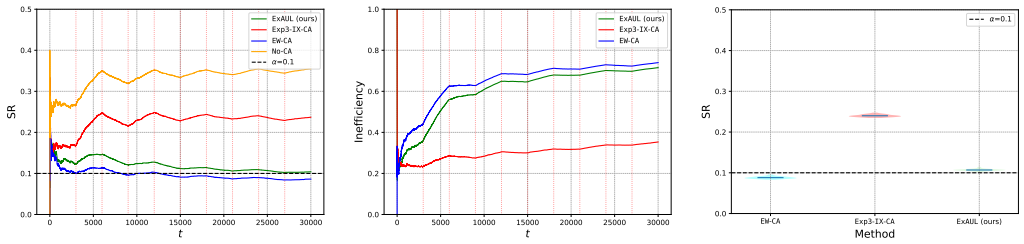
(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 16: Comparison of conformal abstention methods under a distribution-shift environment with LLaMA3.1-8B-Instruct as a generator ($T = 30\text{K}, \alpha = 0.10$), in a distribution shift by alternating between NQ and TriviaQA over time, beginning with NQ. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.

F.2.3 Gradual Shift

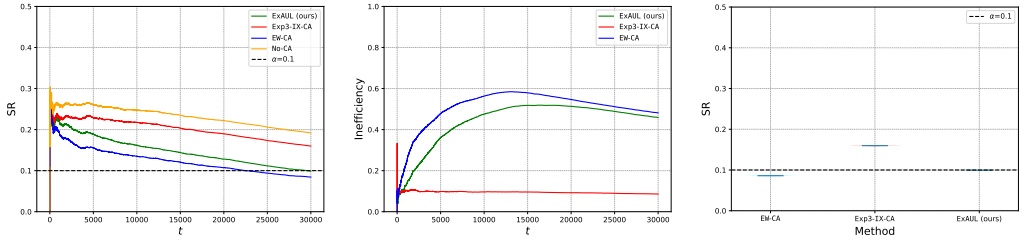
F.2.4 Domain Shift: Open-domain QA to Medical QA

To evaluate a larger domain shift than NQ↔TriviaQA, we consider MedQA-USMLE [18], a medical QA benchmark, and construct streams that shift between an open-domain dataset and MedQA in both directions.



(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 17: Comparison of conformal abstention methods under a distribution-shift environment with LLaMA3.1-8B-Instruct as a generator ($T = 30\text{K}, \alpha = 0.10$), in a distribution shift by alternating between TriviaQA and NQ over time, beginning with TriviaQA. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.



(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 18: Comparison of conformal abstention methods under a distribution-shift environment with GPT-5.4 as a generator ($T = 30\text{K}, \alpha = 0.10$), in a gradual distribution shift from NQ to TriviaQA over time. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.

Despite the substantial domain shift, ExAUL keeps the SR close to α in all four settings (Figures 22–24). Shifting from open-domain QA to MedQA increases inefficiency, while SR control is preserved.

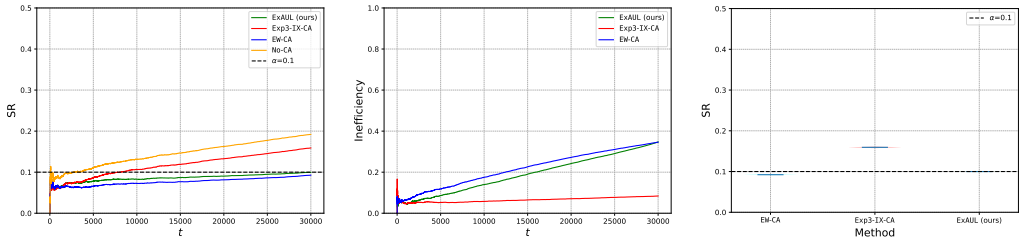
F.3 Imbalanced Streams

The shift experiments above are balanced between the two datasets. To test persistent imbalance, we consider (i) a *steady-state imbalance* setting, where each round is drawn from NQ with probability p_{NQ} and from TriviaQA otherwise, and (ii) an *imbalanced discrete shift*, where a TriviaQA block is followed by an NQ block whose lengths are proportional to $(p_{\text{TriviaQA}}, p_{\text{NQ}})$. All experiments use $T = 60\text{K}$ and $\alpha = 0.1$ with GPT-5.4 and Llama-3.1-8B.

Table 4: SR of ExAUL under imbalanced streams ($T = 60\text{K}$, $\alpha = 0.1$). Each column is the mixing ratio $(p_{\text{NQ}}, p_{\text{TriviaQA}})$.

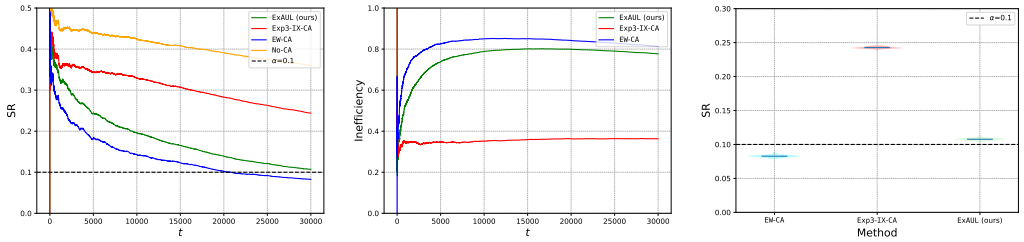
Setting	Model	(0.2,0.8)	(0.4,0.6)	(0.6,0.4)	(0.8,0.2)
Steady-state	GPT-5.4	0.070	0.087	0.105	0.138
	Llama-3.1-8B	0.082	0.093	0.108	0.130
Discrete shift (TriviaQA→NQ)	GPT-5.4	0.072	0.083	0.093	0.111
	Llama-3.1-8B	0.084	0.094	0.107	0.124

In both settings, the SR stays near α for balanced and TriviaQA-dominated streams and rises gradually as NQ dominates (Table 4). This reflects the difference in dataset difficulty rather than a failure of the learner: the per-dataset SR is roughly constant across ratios (about 0.16–0.21 on NQ and 0.05–0.07 on TriviaQA), so the overall SR tracks the share of NQ in the stream, and the minority dataset shows no additional degradation. Thus ExAUL degrades gracefully rather than failing abruptly under persistent imbalance.



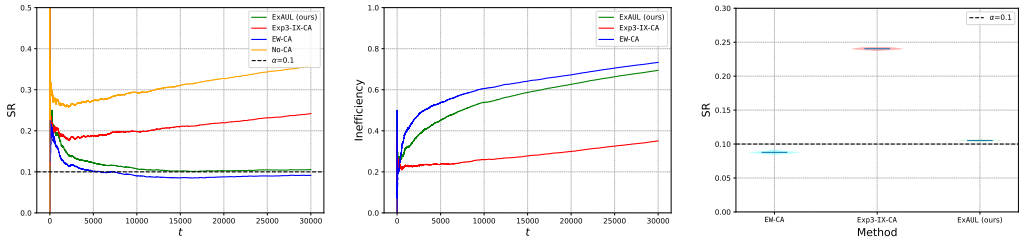
(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 19: Comparison of conformal abstention methods under a gradual distribution-shift environment with GPT-5.4 as a generator ($T = 30\text{K}, \alpha = 0.10$), from TriviaQA to NQ over time. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.



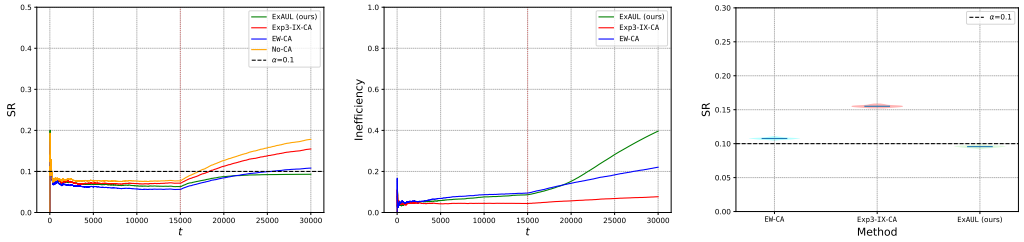
(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 20: Comparison of conformal abstention methods under a distribution-shift environment with LLaMA3.1-8B-Instruct as a generator ($T = 30\text{K}, \alpha = 0.10$), in a gradual distribution shift from NQ to TriviaQA over time. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.



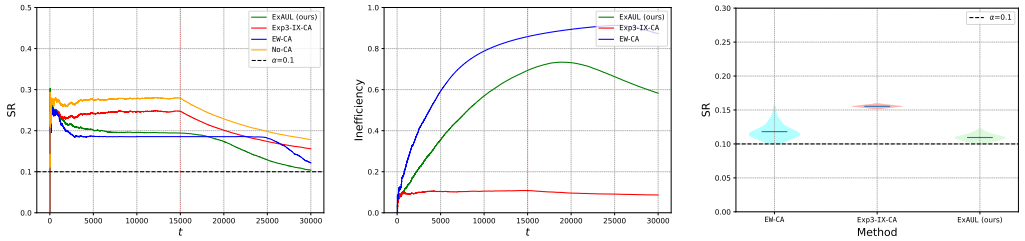
(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 21: Comparison of conformal abstention methods under a distribution-shift environment with LLaMA3.1-8B-Instruct as a generator ($T = 30\text{K}, \alpha = 0.10$), in a gradual distribution shift from TriviaQA to NQ over time. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.



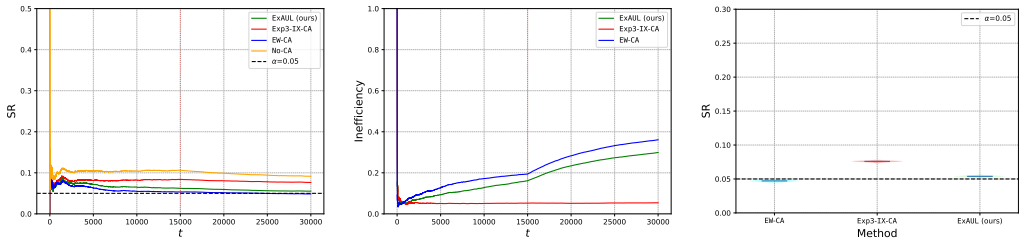
(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 22: Comparison of conformal abstention methods under a distribution-shift environment with GPT-5.4 as a generator ($T = 30\text{K}, \alpha = 0.1$), in a single distribution shift from MedQA to NQ over time. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.



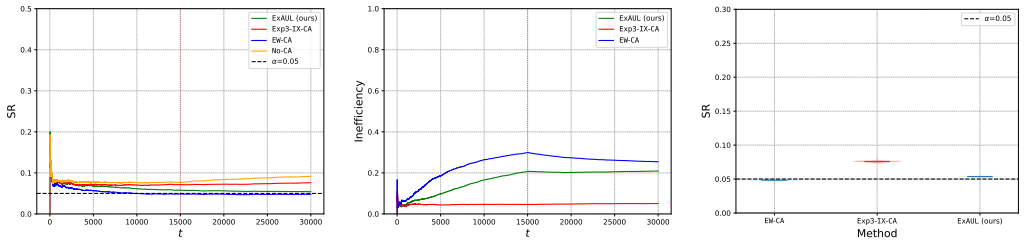
(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 23: Comparison of conformal abstention methods under a distribution-shift environment with GPT-5.4 as a generator ($T = 30\text{K}, \alpha = 0.1$), in a single distribution shift from NQ to MedQA over time. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.



(a) SR over time (b) Inefficiency over time (c) SR distribution

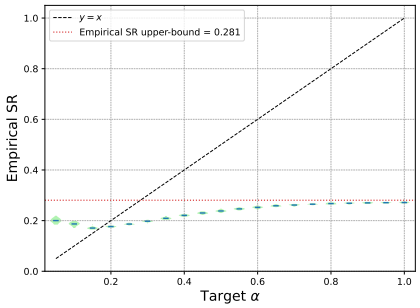
Figure 24: Comparison of conformal abstention methods under a distribution-shift environment with GPT-5.4 as a generator ($T = 30\text{K}, \alpha = 0.05$), in a single distribution shift from TriviaQA to MedQA over time. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.



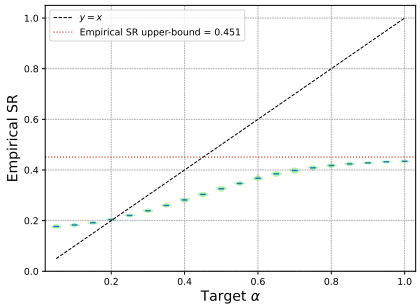
(a) SR over time (b) Inefficiency over time (c) SR distribution

Figure 25: Comparison of conformal abstention methods under a distribution-shift environment with GPT-5.4 as a generator ($T = 30\text{K}, \alpha = 0.05$), in a single distribution shift from MedQA to TriviaQA over time. The violin plots are drawn with randomly chosen 30K samples with 100 random trials.

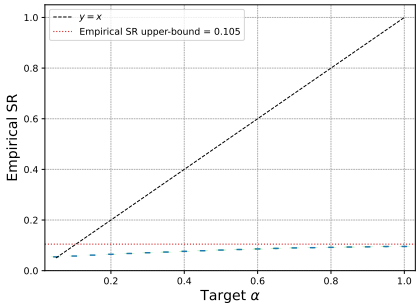
F.4 Analysis on Time Horizon and Desired SR Parameters



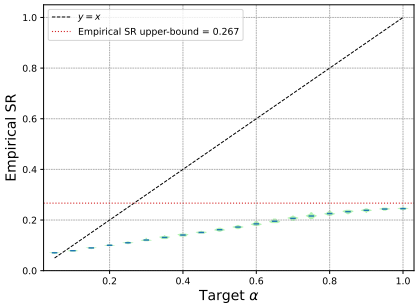
(a) NQ dataset along with GPT-5.4



(b) NQ dataset along with LLaMA3.1-8B

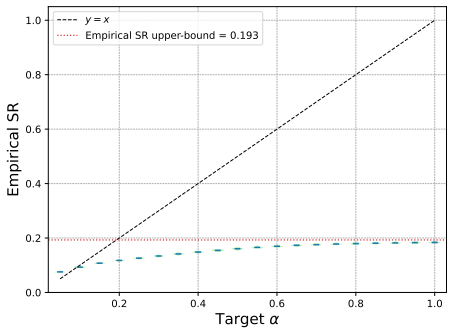


(c) TriviaQA dataset along with GPT-5.4

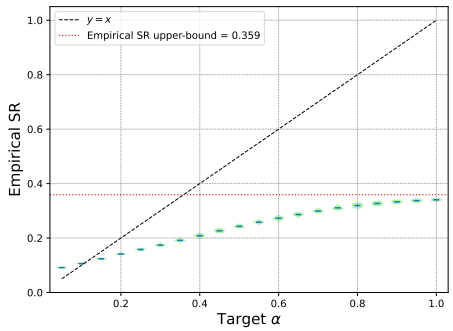


(d) TriviaQA dataset along with LLaMA3.1-8B

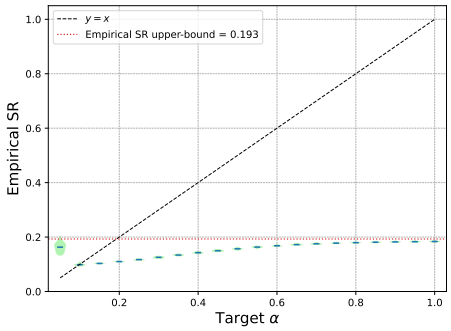
Figure 26: The violin plots for the ablation study over varying α on datasets under stochastic environments. The violin plots are drawn with randomly chosen 30K samples over 30 random trials. Here, converged SRs mostly go along with target risk α (*i.e.,* violin plots are below of the diagonal line), which empirically demonstrates that our algorithm finds efficient conformal abstainers with SR guarantees. Note that $\mathbf{Err}_T$ refers to the error rate on a given dataset, *i.e.,* $\mathbf{Err}_T := \sum_{t=1}^{T} \mathbb{1}\left(G(\mathbf{x}_t) \neq_E \mathbf{y}_t\right)/T$ and the SRs of conformal abstainers should be lower than $\mathbf{Err}_T$ to be meaningful in controlling the SRs. In (a), the SR exceeds α for small α due to the poor calibration of $f_{\texttt{con}}$, not a failure of the guarantee. For NQ with GPT-5.4, 44.5% of the samples have the maximum score $f_{\texttt{con}} \approx 1$ (the largest floating-point value below 1, as $f_t \in [0, 1)$), and 14.8% of them are incorrect. As no threshold can separate these tied samples, every $\tau < 1$ answers all of them, so for α below 14.8%, `ExAUL` is driven toward the always-abstaining threshold $\tau = 1$. The SR violation $\mathcal{V}_T$ in (1), which our theory bounds, then remains controlled since abstention incurs no violation, whereas the empirical $\mathbf{SR}_T$ is a ratio over only a few answered samples and thus becomes large. This is consistent with our SR bound in Section 3.4, which depends on $1 - \mathbf{Ineff}_T$, and thus becomes vacuous when `ExAUL` rarely answers (see Appendix I.1 for further discussion).



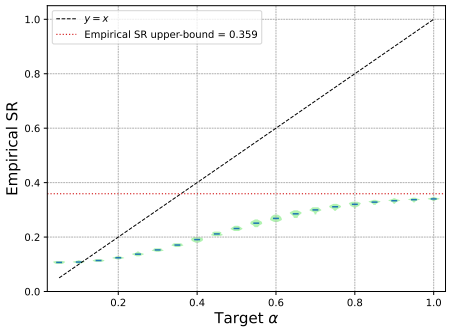
(a) TriviaQA shifts to NQ with GPT-5.4



(b) TriviaQA shifts to NQ with LLaMA3.1-8B

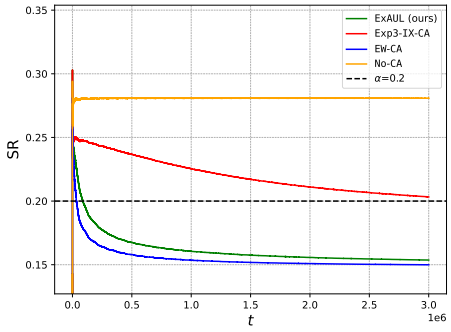


(c) NQ shifts to TriviaQA with GPT-5.4

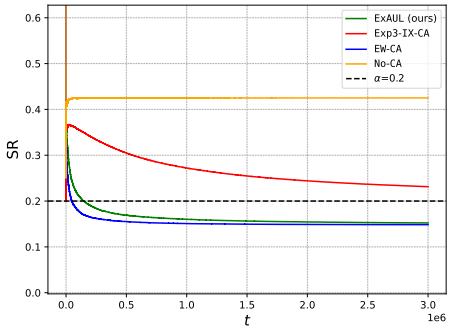


(d) NQ shifts to TriviaQA with LLaMA3.1-8B

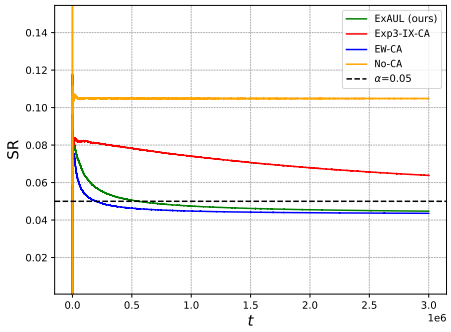
Figure 27: The violin plots for the ablation study over varying α on datasets under distribution-shift environments. The violin plots are drawn with randomly chosen dataset with single distribution shift between 15K-sized TriviaQA and 15K-sized NQ, over 30 random trials. Here, converged SRs mostly go along with target risk α (*i.e.,* violin plots are below the diagonal line), which empirically demonstrates that our algorithm finds efficient conformal abstainers with SR guarantees. Note that $\mathbf{Err}_T$ refers to the error rate on a given dataset, *i.e.,* $\mathbf{Err}_T := \sum_{t=1}^{T} \mathbb{1}\left(G(\mathbf{x}_t) \neq_E \mathbf{y}_t\right)/T$ and the SRs of conformal abstainers should be lower than $\mathbf{Err}_T$ to be meaningful in controlling the SRs. The SR above α for small α in (c) has the same cause as in Figure 26(a): the stream starts with NQ with GPT-5.4, where many samples tie at the maximum score of $f_{\texttt{con}}$, so `ExAUL` is driven toward always abstaining and $\mathbf{SR}_T$ is computed over only a few answered samples, while $\mathcal{V}_T$ remains controlled.



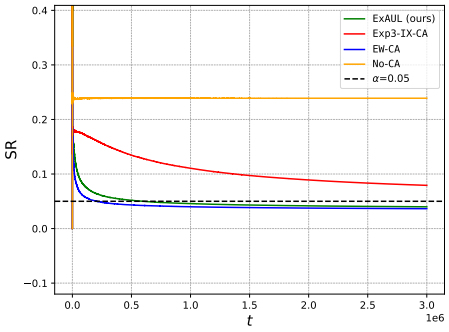
(a) NQ dataset along with GPT-5.4



(b) NQ dataset along with LLaMA3.1-8B

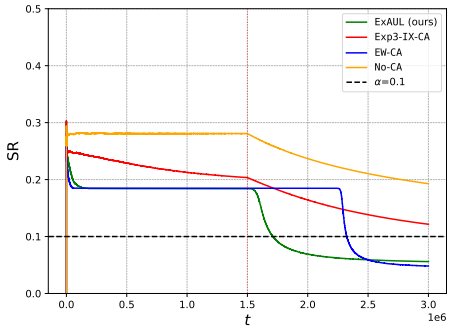


(c) TriviaQA dataset along with GPT-5.4

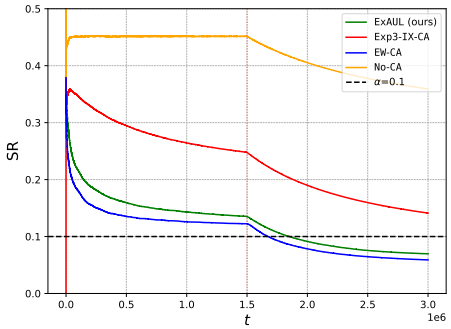


(d) TriviaQA dataset along with LLaMA3.1-8B

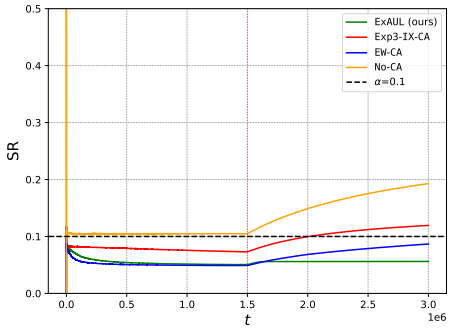
Figure 28: Additional $\mathbf{SR}_t$ comparisons of conformal abstention methods under stochastic environments with a longer time horizon. Here, we repeat the datasets to evaluate whether `Exp3-IX-CA` converges over a long horizon ($T = 3,000$K). As shown in the plots, `Exp3-IX-CA` requires a long horizon T to achieve reliable $\mathbf{SR}_t$ control.



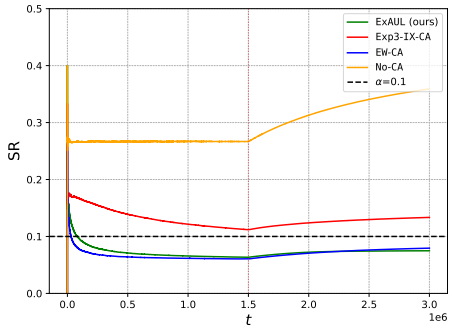
(a) NQ dataset shift to TriviaQA dataset along with GPT-5.4



(b) NQ dataset shift to TriviaQA dataset along with LLaMA3.1-8B



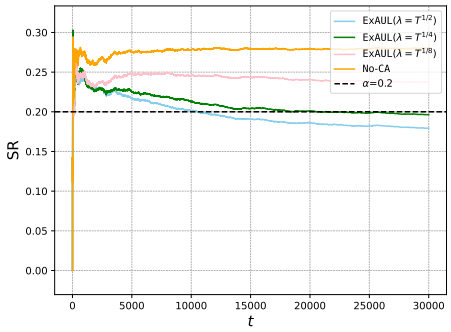
(c) TriviaQA dataset shift to NQ dataset along with GPT-5.4



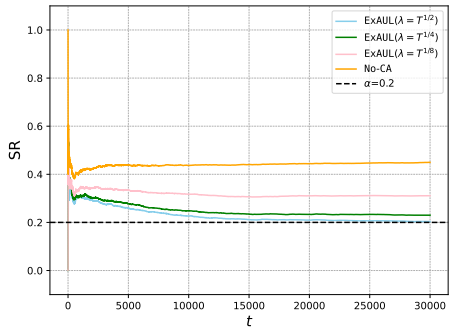
(d) TriviaQA dataset shift to NQ dataset along with LLaMA3.1-8B

Figure 29: Additional $\mathbf{SR}_t$ comparison of conformal abstention methods under distribution-shift environments with a longer time horizon. Here, we repeat the datasets to evaluate whether `Exp3-IX-CA` maintains stable control of $\mathbf{SR}_t$ over a long horizon. We consider a single distribution shift between two $1,500$K-sized datasets (NQ, triviaQA) denoted in a dotted vertical line. As shown in the plots, `Exp3-IX-CA` requires a long horizon T to achieve reliable $\mathbf{SR}_t$ control. In (a), the SR converges to a value above α during the first phase. This is due to the poor calibration of $f_{\texttt{con}}$ for GPT-5.4 on NQ, which leads `EW-CA` and `ExAUL` to always abstain (see Appendix I.1 and Figure 26).

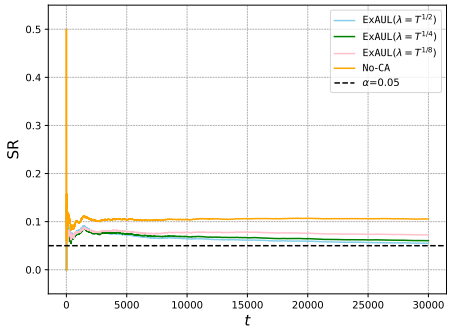
F.5 Analysis on Varying λ



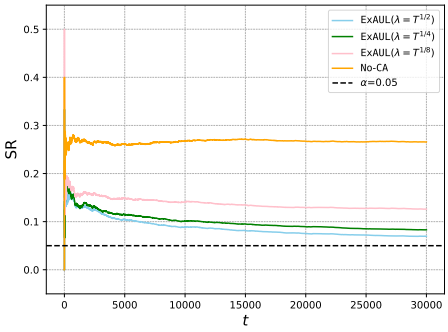
(a) NQ dataset along with GPT-5.4



(b) NQ dataset along with LLaMA3.1-8B

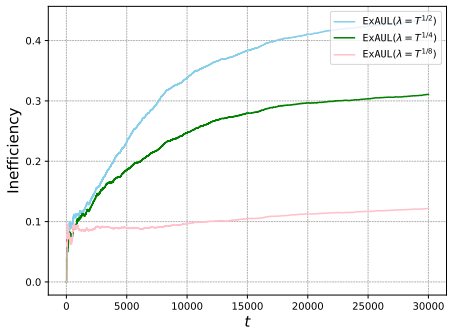


(c) TriviaQA dataset along with GPT-5.4

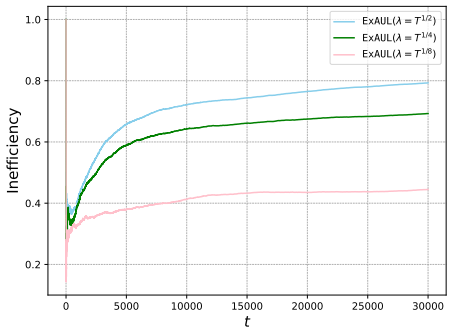


(d) TriviaQA dataset along with LLaMA3.1-8B

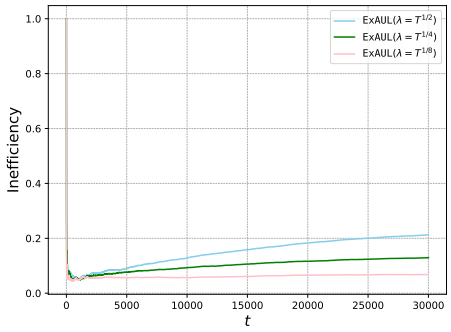
Figure 30: Additional $\mathbf{SR}_t$ comparisons of conformal abstention methods under stochastic environments with varying λ for $T = 30\text{K}$ datasets.



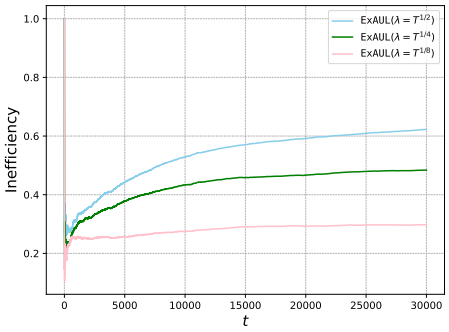
(a) NQ dataset along with GPT-5.4



(b) NQ dataset along with LLaMA3.1-8B



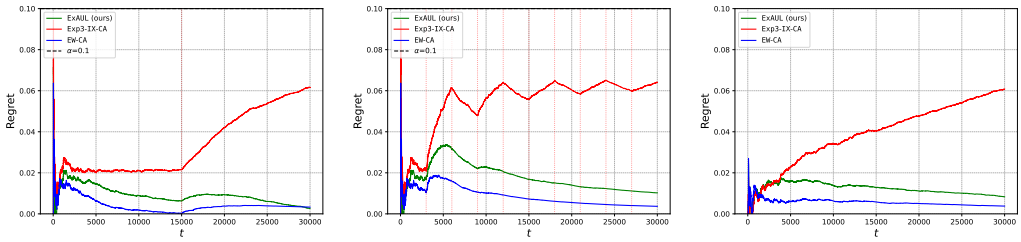
(c) TriviaQA dataset along with GPT-5.4



(d) TriviaQA dataset along with LLaMA3.1-8B

Figure 31: Additional $\mathbf{Ineff}_t$ comparisons of conformal abstention methods under stochastic environments with varying λ for $T = 30\text{K}$ datasets.

F.6 Analysis on the Average Regret in Distribution-shift Environments



(a) Regret in single shift (b) Regret in alternating shift (c) Regret in gradual shift

Figure 32: Comparison of normalized regret trends during training. Similar to SR, regret also shows an increase when a distribution shift occurs. The experimental setups for (a), (b), and (c) correspond to those in Figure 4, 15, and 19, respectively.

F.7 Analysis on Imperfect Feedback

While our framework provides a guarantee for the given feedback, we examine for completeness how ExAUL behaves when the feedback is perturbed. In all experiments below, ExAUL learns from the perturbed feedback, while we report the SR against the original labels from GPT-5.4-nano.

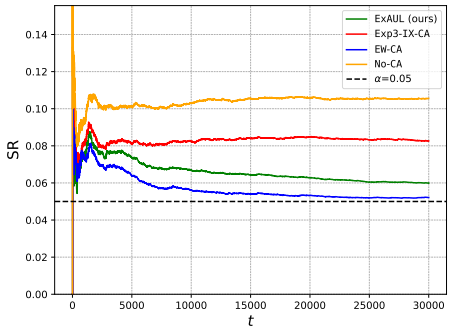
Weak supervision. We first replace the feedback with weaker proxies, namely thresholded ROUGE-L scores and randomly flipped labels (Figures 33 and 34).

Label noise. We further flip the feedback with symmetric noise, false-negative noise (correct-to-incorrect flips), and false-positive noise (incorrect-to-correct flips) at rates of 5%, 10%, and 20%, on NQ with GPT-5.4 and Llama-3.1-8B ($T = 30$K, $\alpha = 0.2$, $f_{\texttt{con}}$).

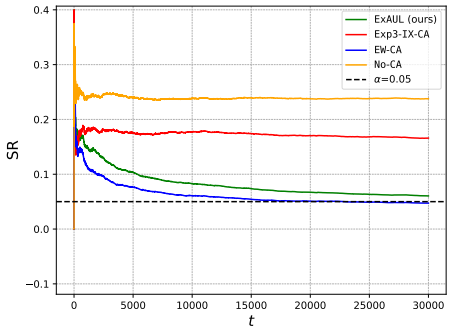
Table 5: Empirical SR under imperfect feedback.

		GPT-5.4			Llama-3.1-8B		
Noise	Rate	Exp3-IX-CA	ExAUL	EW-CA	Exp3-IX-CA	ExAUL	EW-CA
No-CA		0.282			0.448		
Clean	–	0.246	**0.181**	0.166	0.344	**0.199**	0.172
Sym.	5%	0.246	**0.179**	0.161	0.345	**0.197**	0.159
	10%	0.246	**0.191**	0.184	0.345	**0.193**	0.151
	20%	0.246	**0.229**	0.214	0.345	**0.197**	0.165
FN	5%	0.245	**0.177**	0.160	0.343	**0.193**	0.155
	10%	0.245	**0.199**	0.195	0.342	**0.187**	0.146
	20%	0.245	**0.224**	0.208	0.340	**0.183**	0.155
FP	5%	0.246	**0.182**	0.168	0.346	**0.204**	0.175
	10%	0.246	**0.185**	0.172	0.348	**0.208**	0.177
	20%	0.247	**0.195**	0.181	0.351	**0.223**	0.193

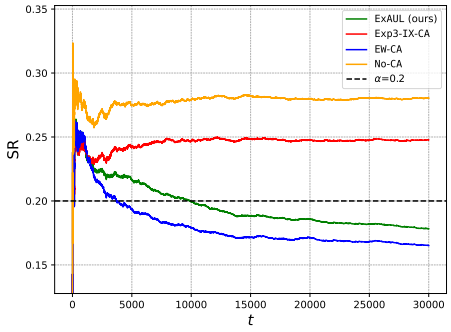
Analysis. As shown in Table 5, the SR of ExAUL exceeds α by at most 0.008 for flip rates up to 10% and by at most 0.029 at 20%, and the most harmful noise type depends on the generator. EW-CA attains a lower SR than ExAUL in every setting, while Exp3-IX-CA does not reach α in any setting. Obtaining reliable feedback from noisy or biased signals, *e.g.,* by denoising or aggregating user feedback, is a separate research direction that is complementary to ours, since any improved feedback can be directly used by ExAUL .



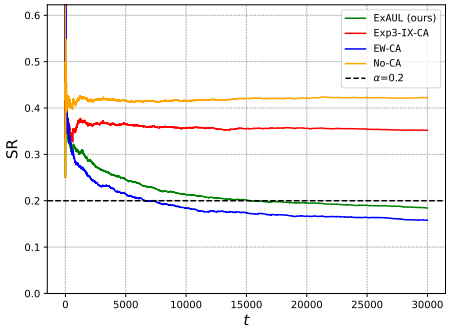
(a) TriviaQA dataset along with GPT-5.4



(b) TriviaQA dataset along with LLaMA3.1-8B

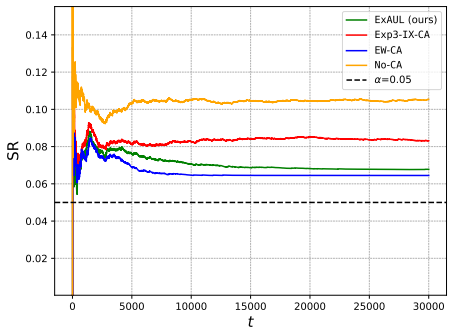


(c) NQ dataset along with GPT-5.4

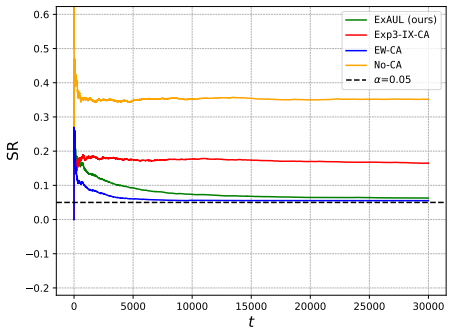


(d) NQ dataset along with LLaMA3.1-8B

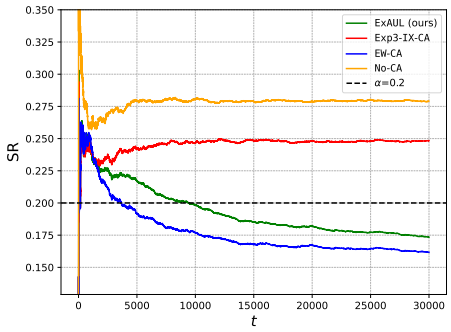
Figure 33: Comparison of conformal abstention methods under an imperfect supervision environment on the NQ ($T = 30K$, $\alpha = 0.2$) and TriviaQA ($T = 30K, \alpha = 0.05$) along with GPT-5.4 and LLaMA3.1-8B, same as Figure 3, 6, 7, and 8. Here, "noisy" feedback is generated by randomly flipping 5% of the feedback labels. For a fair comparison, we use probability $0.05/(2 * \texttt{base error rate})$ for incorrect labels and $0.05/[2(1 - \texttt{base error rate})]$ for correct ones. Similar to the ROUGE-L experiment (Figure 34), `ExAUL` demonstrates robustness, successfully learning to control the SR even under conditions of label noise.



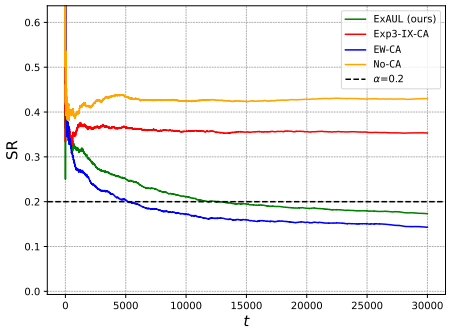
(a) TriviaQA dataset along with GPT-5.4



(b) TriviaQA dataset along with LLaMA3.1-8B



(c) NQ dataset along with GPT-5.4



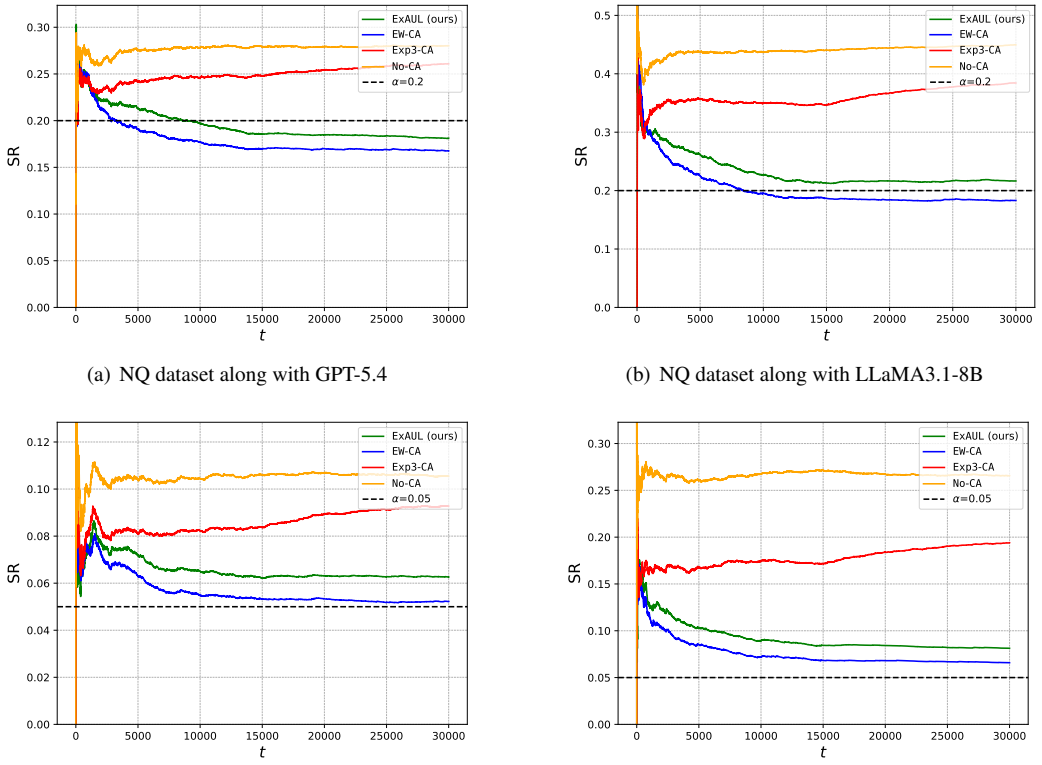
(d) NQ dataset along with LLaMA3.1-8B

Figure 34: Comparison of conformal abstention methods under an imperfect supervision environment on the NQ ($T = 30K$, $\alpha = 0.2$) and TriviaQA ($T = 30K, \alpha = 0.05$) along with GPT-5.4 and LLaMA3.1-8B, same as Figure 3, 6, 7, and 8. Here, "noisy" feedback is simulated using thresholded ROUGE-L scores as the supervision signal. For a fair comparison, we use different thresholds for each (dataset, model) pair to maintain the original error rate. The plots demonstrate that ExAUL maintains its performance advantage, successfully controlling the SR near the target α despite the weaker, heuristic-based feedback.

F.8 Analysis on scoring function

Our theoretical guarantee is formulated over a class of thresholds applied to a given confidence score and does not require the score to be calibrated or informative. To empirically evaluate the robustness of our method to changes in score quality, we consider two settings: switching between different scoring functions and corrupting confidence scores with random noise.

Robustness to Different Scoring Functions We first investigate whether our method remains effective when the scoring function changes during deployment. Specifically, we replace the original scoring function $f_{\texttt{con}}$ with $f_{\texttt{std}}$ at the midpoint of the stream ($t = T/2$), while keeping the generator, feedback, and evaluation labels unchanged.



(a) NQ dataset along with GPT-5.4

(b) NQ dataset along with LLaMA3.1-8B

(c) TriviaQA dataset along with GPT-5.4

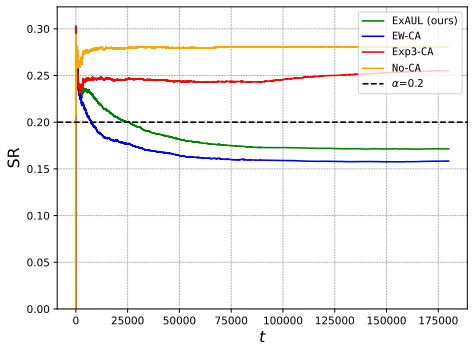
(d) TriviaQA dataset along with LLaMA3.1-8B

Figure 35: Comparison of $\mathbf{SR}_t$ across conformal abstention methods when the scoring function changes from $f_{\texttt{con}}$ to $f_{\texttt{std}}$ over $T = 30\text{K}$ datasets.

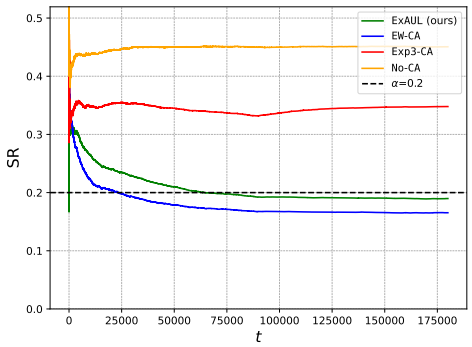
As shown in Fig. 35, ExAUL maintains its SR relatively close to the target level α across different datasets and generators, despite the change in scoring function. Although the empirical SR does not always remain below α for $T = 30\text{K}$, Fig. 36 shows that extending the horizon to $T = 180\text{K}$ further reduces the SR, bringing it closer to the target level. These results suggest that ExAUL can adapt to changes in the scoring function and progressively approach the target risk level over longer horizons.

Robustness to Uncalibrated Confidence Scores We further investigate robustness to uninformative confidence scores by randomly corrupting the original scores. At each round, independently with probability ε, the original confidence score is replaced by a random value $u_t \sim \mathcal{U}(0, 1)$, independent of the correctness of the generated answer:

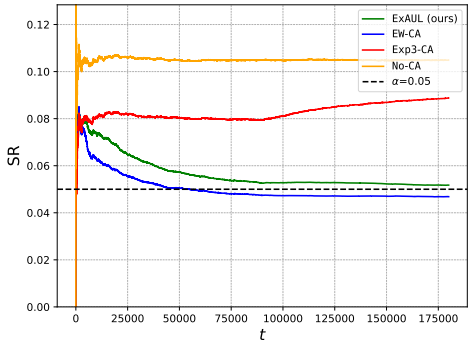
$$\tilde{s}_t = \begin{cases} u_t, & \text{with probability } \varepsilon, \\ f_{\texttt{con}}(x_t), & \text{otherwise.} \end{cases}$$



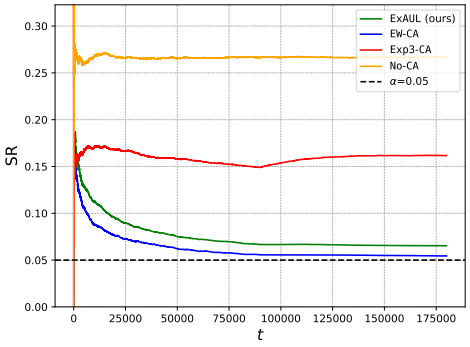
(a) NQ dataset along with GPT-5.4



(b) NQ dataset along with LLaMA3.1-8B



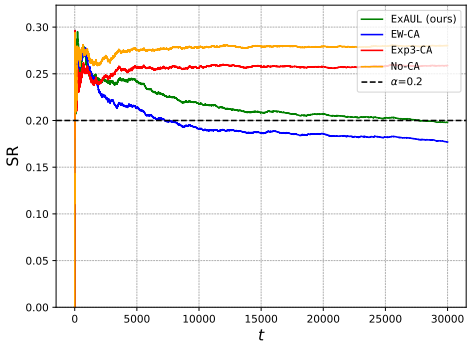
(c) TriviaQA dataset along with GPT-5.4



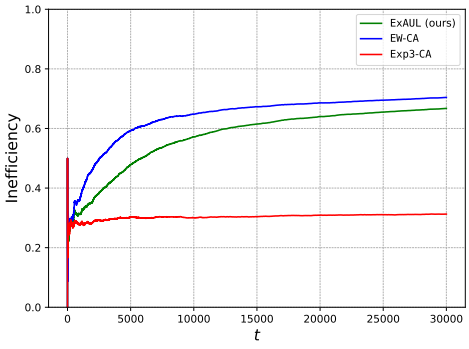
(d) TriviaQA dataset along with LLaMA3.1-8B

Figure 36: Comparison of $\mathbf{SR}_t$ across conformal abstention methods when the scoring function changes from $f_{\texttt{con}}$ to $f_{\texttt{std}}$ over $T = 180\text{K}$ datasets.

Here, $\varepsilon = 0$ recovers the original score stream, while $\varepsilon = 1$ makes the scores entirely non-informative. We set $\varepsilon = 0.5$ while keeping the generator, feedback, and evaluation labels unchanged.



(a) SR over time



(b) Inefficiency over time

Figure 37: Robustness to confidence-score corruption on NQ ($T = 30\text{K}, \alpha = 0.20$) with GPT-5.4.

As shown in Fig. 37, `ExAUL` maintains its SR near the target level α despite the corruption of confidence scores. Moreover, compared with the original setting using only $f_{\texttt{con}}$ in Fig. 7, we observe a reduction in efficiency, indicating a shift in the risk–efficiency trade-off. These results suggest that `ExAUL` can remain effective even when the confidence scores become less informative.

Table 6: Claim-level SR and inefficiency on ASQA with GPT-5.4-nano ($\alpha = 0.3$), as mean $\pm$ standard deviation over five learner seeds.

Method	SR	Inefficiency
`No-CA`	0.4776	0.0000
`Exp3-IX-CA`	0.3564 ± 0.0020	0.4148 ± 0.0026
`ExAUL`	$\mathbf{0.2535 \pm 0.0043}$	0.6816 ± 0.0125
`EW-CA`	0.2110 ± 0.0018	0.7375 ± 0.0020

F.9 Long-Form Generation

Recent long-form methods decompose a response into atomic claims and estimate confidence or verify correctness per claim [33, 47, 51, 52, 17]. Given such claim-level scores and feedback, `ExAUL` applies without modification by treating each claim as one online round; independence between claims is not required, since our protocol permits an adaptive adversary (Section 2).

Setup. On ASQA [43], GPT-5.4-nano generates 3–5 sentence responses, extracts atomic claims, and judges their factuality, and the score is $P(\mathrm{True})$ from the first-token probabilities of "True" and "False" [51]. We exclude claims judged unverifiable, since absence from the reference does not imply incorrectness, which yields $3{,}528$ claims from 987 responses. We set $e_t = \mathbb{1}(\text{claim is contradicted})$, $\alpha = 0.3$, and $|\mathcal{H}| = 1\mathrm{K}$, with five epochs over the claim stream ($T \approx 18\mathrm{K}$).

Results. As shown in Table 6, `ExAUL` keeps the claim-level SR below α while retaining 31.8% of the claims, whereas `Exp3-IX-CA` does not reach the target, and `EW-CA` with full feedback abstains even more often. This check only shows that `ExAUL` operates at the claim level; a complete long-form system would also require reliable claim extraction, verification, and reconstruction, which we leave to future work.

G Theorem on Our Main Regret Bound

Theorem 4. *Let $\ell_t(\cdot) \in [0,1]$ of the form (4). For any $T \in \mathbb{N}$, $\alpha \in (0,1)$, and finite hypotheses $\mathcal{H}$, Algorithm 1 with $\eta = 2\gamma = \sqrt{\frac{\ln|\mathcal{H}|}{T}}$ provides the following regret bound with probability at least $1-\delta$:*

$$\boldsymbol{Reg}_T \le 4\sqrt{T \ln|\mathcal{H}|} + \left(1 + \sqrt{\frac{T}{\ln|\mathcal{H}|}}\right) \ln\frac{2}{\delta}.$$

H Proofs

H.1 A Proof of Theorem 3

We follow the standard proof [12], followed by extending it for a loss range with $[0, \ell_{\max}]$.

Let the cumulative loss of the i-th hypothesis $\tau_i \in \mathcal{H}$ be $L_t^i \coloneqq \sum_{s=1}^t \ell_s(\tau_i)$, $W_t \coloneqq \sum_{i=1}^{|\mathcal{H}|} \exp\left\{-\eta L_t^i\right\}$.

The upper bound of $\ln \frac{W_t}{W_{t-1}}$. We have the following upper bound of $\ln \frac{W_t}{W_{t-1}}$:

$$\begin{aligned}
\ln \frac{W_t}{W_{t-1}} &= \ln \frac{\sum_{i=1}^{|\mathcal{H}|} \exp\left\{-\eta L_t^i\right\}}{\sum_{i=1}^{|\mathcal{H}|} \exp\{-\eta L_{t-1}^i\}} \\
&= \ln \sum_{i=1}^{|\mathcal{H}|} \exp\left\{-\eta \ell_t(\tau_i)\right\} \frac{\exp\{-\eta L_{t-1}^i\}}{\sum_{i=1}^{|\mathcal{H}|} \exp\{-\eta L_{t-1}^i\}} \\
&= \ln \mathbb{E}_{\tau_t \sim p_t} \exp\left\{-\eta \ell_t(\tau_t)\right\} \\
&\leq -\eta \mathbb{E}_{\tau_t \sim p_t} \ell_t(\tau_t) + \frac{\eta^2 \ell_{\max}^2}{8}, \qquad (15)
\end{aligned}$$

where (15) holds by the Hoeffding's lemma and the fact that $\ell_t(\tau_i) \in [0, \ell_{\max}]$. Thus, we have

$$\ln \frac{W_T}{W_0} = \sum_{t=1}^T \ln \frac{W_t}{W_{t-1}} \leq -\eta \sum_{t=1}^T \mathbb{E}_{\tau_t \sim p_t} \ell_t(\tau_t) + \frac{\eta^2 \ell_{\max}^2 T}{8}.$$

The lower bound of $\ln \frac{W_t}{W_{t-1}}$. We have the lower bound of $\ln \frac{W_t}{W_{t-1}}$ as follows:

$$\begin{aligned}
\ln \frac{W_T}{W_0} = \ln \sum_{i=1}^{|\mathcal{H}|} \exp\{-\eta L_T^i\} - \ln |\mathcal{H}| &\geq \ln \left(\max_{i \in \{1,\ldots,|\mathcal{H}|\}} \exp\{-\eta L_T^i\} \right) - \ln |\mathcal{H}| \\
&= -\eta \min_{i \in \{1,\ldots,|\mathcal{H}|\}} L_T^i - \ln |\mathcal{H}| \\
&= -\eta \min_{\tau \in \mathcal{H}} \sum_{t=1}^T \ell_t(\tau) - \ln |\mathcal{H}|.
\end{aligned}$$

Combine the lower and upper bounds. If we combine the lower bound and the upper bound, we obtain the following inequality:

$$\sum_{t=1}^T \mathbb{E}_{\tau_t \sim p_t} \ell_t(\tau_t) - \min_{\tau \in \mathcal{H}} \sum_{t=1}^T \ell_t(\tau) \leq \frac{\eta \ell_{\max}^2 T}{8} + \frac{\ln |\mathcal{H}|}{\eta}.$$

If $\eta = \sqrt{8 \ln |\mathcal{H}| / \ell_{\max}^2 T}$,

$$\sum_{t=1}^T \mathbb{E}_{\tau_t \sim p_t} \ell_t(\tau_t) - \min_{\tau \in \mathcal{H}} \sum_{t=1}^T \ell_t(\tau) \leq \ell_{\max} \sqrt{\frac{T \ln |\mathcal{H}|}{2}}. \qquad (16)$$

Since the guarantee in (16) holds for any sequence of loss functions in `EW`, the bound is also valid against an adaptive adversary. Furthermore, the expected regret is given by $\mathbb{E}[\mathbf{Reg}_T] = \mathbb{E}\left[\sum_{t=1}^T \ell_t(\tau_t) - \min_{\tau \in \mathcal{H}} \sum_{t=1}^T \ell_t(\tau)\right]$, which follows from the tower property of expectation.

High probability bound. Moreover, we have

$$\sum_{t=1}^T \ell_t(\tau_t) - \min_{\tau \in \mathcal{H}} \sum_{t=1}^T \ell_t(\tau) = \left(\sum_{t=1}^T \ell_t(\tau_t) - \sum_{t=1}^T \mathbb{E}_{\tau_t \sim p_t} \ell_t(\tau_t) \right) + \left(\sum_{t=1}^T \mathbb{E}_{\tau_t \sim p_t} \ell_t(\tau_t) - \min_{\tau \in \mathcal{H}} \sum_{t=1}^T \ell_t(\tau) \right).$$

Since $\mathbb{E}_t\left[\ell_t(\tau_t) - \mathbb{E}_{\tau_t\sim p_t}\ell_t(\tau_t)\right] = 0$, which implies that $\sum_{t=1}^{T}\left[\ell_t(\tau_t) - \mathbb{E}_{\tau_t\sim p_t}\ell_t(\tau_t)\right]$ is a martingale, we can use Hoeffding-Azuma inequality

$$\mathbb{P}\left(\sum_{t=1}^{T}\left[\ell_t(\tau_t) - \mathbb{E}_{\tau_t\sim p_t}\ell_t(\tau_t)\right] \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{t=1}^{T} c_t^2}\right)$$

for all $\varepsilon \geq 0$ where $c_t = \ell_{\max}$ since $\ell_t(\tau_t) - \mathbb{E}_{\tau_t\sim p_t}\ell_t(\tau_t) \in [-\mathbb{E}_{\tau_t\sim p_t}\ell_t(\tau_t), \ell_{\max} - \mathbb{E}_{\tau_t\sim p_t}\ell_t(\tau_t)]$.

Thus, if we put $\delta = \exp\left(-\frac{2\varepsilon^2}{T\ell_{\max}^2}\right)$ we get

$$\sum_{t=1}^{T}\ell_t(\tau_t) - \sum_{t=1}^{T}\mathbb{E}_{\tau_t\sim p_t}\ell_t(\tau_t) \leq \ell_{\max}\sqrt{\frac{T\ln\delta^{-1}}{2}},$$

with probability at least $1-\delta$, which suggests that

$$\sum_{t=1}^{T}\ell_t(\tau_t) - \min_{\tau\in\mathcal{H}}\sum_{t=1}^{T}\ell_t(\tau) \leq \frac{\eta\ell_{\max}^2 T}{8} + \frac{\ln|\mathcal{H}|}{\eta} + \ell_{\max}\sqrt{\frac{T\ln\delta^{-1}}{2}}$$

with probability at least $1-\delta$.

If $\eta = \sqrt{8\ln|\mathcal{H}|/\ell_{\max}^2 T}$,

$$\sum_{t=1}^{T}\ell_t(\tau_t) - \min_{\tau\in\mathcal{H}}\sum_{t=1}^{T}\ell_t(\tau) \leq \ell_{\max}\sqrt{\frac{T\ln|\mathcal{H}|}{2}} + \ell_{\max}\sqrt{\frac{T\ln\delta^{-1}}{2}}. \tag{17}$$

H.2 A Proof of Theorem 2

This proof follows standard proof [36] but with a non-trivial loss range $[0, \ell_{\max}]$. Recall that $\ell_t(\tau) \in [0, \ell_{\max}]$ and its biased estimator $\ell_t(\tau \mid \{\tau_t\}) \in [0, \infty)$, as defined in (12).

We first introduce new Lemma 2 and Corollary 1 that establish a high probability bound between the loss estimator and the empirical loss. Corollary 1 provides a bound for a fixed value of τ, and its proof follows directly from the Lemma 2 by applying a union bound. Together, Lemma 2 and Corollary 1 allow us to translate the upper bound on the estimator equation into a high-probability regret bound.

Lemma 2. *Let $\gamma_t \geq 0$ for $t \in \{1, 2, \ldots, T\}$ be a fixed non-increasing sequence and $\alpha_{t,\tau}$ be a real value that satisfies $\alpha_{t,\tau} \leq 2\gamma_t$ for all $t \in \mathbb{N}$ and $\tau \in \mathcal{H}$. Then, we have*

$$\sum_{t=1}^{T}\sum_{\tau\in\mathcal{H}}\alpha_{t,\tau}\left(\ell_t(\tau \mid \{\tau_t\}) - \ell_t(\tau)\right) \leq \ell_{max}\ln\frac{1}{\delta},$$

with probability $1-\delta$.

Proof. First, letting $\beta_t = \frac{2\gamma_t}{\ell_{\max}}$, we have

$$\begin{aligned}
\ell_t(\tau \mid \{\tau_t\}) &= \frac{\ell_t(\tau)}{p_t(\tau)+\gamma_t}\cdot\mathbb{1}(\tau_t = \tau) \\
&= \ell_{\max}\cdot\frac{\ell_t(\tau)/\ell_{\max}}{p_t(\tau)+\gamma_t}\cdot\mathbb{1}(\tau_t = \tau) \\
&\leq \ell_{\max}\cdot\frac{\ell_t(\tau)/\ell_{\max}}{p_t(\tau)+\gamma_t\cdot\ell_t(\tau)/\ell_{\max}}\cdot\mathbb{1}(\tau_t = \tau) \\
&= \frac{\ell_{\max}}{2\gamma_t}\cdot\frac{2\gamma_t\cdot\frac{\ell_t(\tau)}{\ell_{\max}p_t(\tau)}}{1+\gamma_t\cdot\frac{\ell_t(\tau)}{\ell_{\max}p_t(\tau)}}\mathbb{1}(\tau_t = \tau) \\
&\leq \frac{\ell_{\max}}{2\gamma_t}\cdot\ln\left(1+2\gamma_t\cdot\frac{\ell_t(\tau)}{\ell_{\max}p_t(\tau)}\mathbb{1}(\tau_t = \tau)\right) \qquad (18) \\
&= \frac{1}{\beta_t}\cdot\ln\left(1+\beta_t\frac{\ell_t(\tau)}{p_t(\tau)}\mathbb{1}(\tau_t = \tau)\right) \qquad (19)
\end{aligned}$$

where the inequality (18) holds by $\frac{x}{1+x/2} \leq \ln(1+x)$ for $x \geq 0$ and the range $\{0,1\}$ of the indicator function.

Second, we explicitly define the following auxiliary quantities:

$$\tilde{S}_t = \sum_{\tau \in \mathcal{H}} \frac{\alpha_{t,\tau}}{\ell_{\max}} \ell_t(\tau \mid \{\tau_t\}) \quad \text{and} \quad S_t = \sum_{\tau \in \mathcal{H}} \frac{\alpha_{t,\tau}}{\ell_{\max}} \ell_t(\tau).$$

Denoting by $\mathbb{E}_t$ the conditional expectation with respect to τ_t, given the past learner's choices $\tau_1, \ldots, \tau_{t-1}$, we have the following relation between $\tilde{S}_t$ and S_t:

$$\begin{aligned}
\mathbb{E}_t\left[\exp \tilde{S}_t\right] &\leq \mathbb{E}_t\left[\exp\left(\sum_{\tau \in \mathcal{H}} \frac{\alpha_{t,\tau}}{\ell_{\max}\beta_t} \cdot \ln\left(1 + \beta_t \cdot \frac{\ell_t(\tau)}{p_t(\tau)} \mathbb{1}(\tau_t = \tau)\right)\right)\right] && (20)\\
&= \mathbb{E}_t\left[\prod_{\tau \in \mathcal{H}} \left(1 + \beta_t \frac{\ell_t(\tau)}{p_t(\tau)} \mathbb{1}(\tau_t = \tau)\right)^{\frac{\alpha_{t,\tau}}{\ell_{\max}\beta_t}}\right] \\
&\leq \mathbb{E}_t\left[1 + \sum_{\tau \in \mathcal{H}} \frac{\alpha_{t,\tau}}{\ell_{\max}} \cdot \frac{\ell_t(\tau)}{p_t(\tau)} \cdot \mathbb{1}(\tau_t = \tau)\right] && (21)\\
&\leq 1 + \sum_{\tau \in \mathcal{H}} \frac{\alpha_{t,\tau}}{\ell_{\max}} \cdot \ell_t(\tau) \\
&\leq \exp\left(\sum_{\tau \in \mathcal{H}} \frac{\alpha_{t,\tau}}{\ell_{\max}} \cdot \ell_t(\tau)\right) = \exp\left(S_t\right). && (22)
\end{aligned}$$

Here, the inequality (20) follows directly from (19). Subsequently, (21) is obtained by applying the Bernoulli inequality $(1+z)^a \leq (1+az)$ which is valid for $z \geq 0$ and $0 \leq a \leq 1$ and the condition of $\alpha_t \leq 2\gamma_t$, followed by the property of the indicator function.

Finally we define:

$$W_t := \exp\left(\sum_{i=1}^{t} \left(\tilde{S}_i - S_i\right)\right).$$

This immediately implies that $W_t = W_{t-1} \cdot \exp\left(\tilde{S}_t - S_t\right)$. Taking the conditional expectation with respect to τ_t, we obtain

$$\begin{aligned}
\mathbb{E}_t[W_t] &= \mathbb{E}_t\left[W_{t-1} \cdot \exp\left(\tilde{S}_t - S_t\right)\right] \\
&= W_{t-1}\mathbb{E}_t\left[\exp\left(\tilde{S}_t - S_t\right)\right] \\
&= W_{t-1}\exp\left(-S_t\right)\mathbb{E}_t\left[\exp\left(\tilde{S}_t\right)\right] && (23)\\
&\leq W_{t-1}\exp\left(-S_t\right)\exp\left(S_t\right) && (24)\\
&\leq W_{t-1}, && (25)
\end{aligned}$$

where (23) W_{t-1} and S_t are dependent on the history up to time $t-1$, we pull W_{t-1} and $\exp(-S_t)$ outside of the conditional expectation. Also, (24) holds due to (22).

From (25), we have $\mathbb{E}_T[W_T] \leq \mathbb{E}_{T-1}[W_{T-1}] \leq \cdots \leq \mathbb{E}_0[W_0]$. As $W_0 = 1$, we have $\mathbb{E}_T[W_T] \leq 1$. Then, by this and the Markov's inequality, we have

$$\mathbb{P}\left(\sum_{t=1}^{T}(\tilde{S}_t - S_t) > \ln\frac{1}{\delta}\right) = \mathbb{P}\left(W_T > \frac{1}{\delta}\right) \leq \delta\mathbb{E}_T[W_T] \leq \delta.$$

Recall that $\ln W_T = \sum_{t=1}^{T}\left(\tilde{S}_t - S_t\right) = \sum_{t=1}^{T}\sum_{\tau \in \mathcal{H}} \frac{\alpha_{t,\tau}}{\ell_{\max}}\left(\ell_t(\tau \mid \{\tau_t\}) - \ell_t(\tau)\right)$ then we have

$$\sum_{t=1}^{T}\sum_{\tau \in \mathcal{H}} \alpha_{t,\tau}\left(\ell_t(\tau \mid \{\tau_t\}) - \ell_t(\tau)\right) \leq \ell_{\max}\ln\frac{1}{\delta}$$

with probability at least $1-\delta$, which completes the proof. □

Corollary 1. *Let $\gamma_t \geq \gamma > 0$ for all t. Simultaneously for all $\tau \in \mathcal{H}$, we have*

$$\sum_{t=1}^{T} \ell_t(\tau \mid \{\tau_t\}) - \sum_{t=1}^{T} \ell_t(\tau) \leq \frac{\ell_{max}}{2\gamma} \ln \frac{|\mathcal{H}|}{\delta} \tag{26}$$

with probability $1 - \delta$.

Proof. For each $\bar{\tau} \in \mathcal{H}$, let $\alpha_{t,\tau} \coloneqq 2\gamma \mathbb{1}(\tau = \bar{\tau}) \leq 2\gamma_t$, which satisfies the condition for Lemma 2. Then, due to Lemma 2, we have

$$\begin{aligned}
\ell_{\max} \ln \frac{|\mathcal{H}|}{\delta} &\geq \sum_{t=1}^{T} \sum_{\tau \in \mathcal{H}} \alpha_{t,\tau} \left(\ell_t(\tau \mid \{\tau_t\}) - \ell_t(\tau) \right) \\
&= \sum_{t=1}^{T} \sum_{\tau \in \mathcal{H}} 2\gamma \mathbb{1}(\tau = \bar{\tau}) \left(\ell_t(\tau \mid \{\tau_t\}) - \ell_t(\tau) \right) \\
&= \sum_{t=1}^{T} 2\gamma \left(\ell_t(\bar{\tau} \mid \{\tau_t\}) - \ell_t(\bar{\tau}) \right)
\end{aligned}$$

with probability at least $1 - \frac{\delta}{|\mathcal{H}|}$. By taking the union bound for all $\bar{\tau} \in \mathcal{H}$, we complete the proof. □

Then, we provide the main proof.

First step. We split the expected loss estimator into two logarithmic terms Eq. (3.7) in [6]: a variability term and the log of an exponential expectation. We will rewrite each term in the following steps.

$$\begin{aligned}
\mathbb{E}_{\tau \sim p_t} \ell(\tau \mid \{\tau_t\}) = \frac{1}{\eta_t} \underbrace{\ln \mathbb{E}_{\tau \sim p_t} \exp\Big(- \eta_t \Big(\ell_t(\tau \mid \{\tau_t\}) - \mathbb{E}_{\bar{\tau} \sim p_t} \ell_t(\bar{\tau} \mid \{\tau_t\}) \Big) \Big)}_{\text{variability term}} \\
- \frac{1}{\eta_t} \underbrace{\ln \mathbb{E}_{\tau \sim p_t} \exp\Big(- \eta_t \ell_t(\tau \mid \{\tau_t\}) \Big)}_{\text{log of an exponential expectation term}}.
\end{aligned} \tag{27}$$

Second step. We provide an upper bound on the variability term.

$$\begin{aligned}
&\ln \mathbb{E}_{\tau} \exp\Big(- \eta_t \left(\ell_t(\tau \mid \{\tau_t\}) - \mathbb{E}_{\bar{\tau}} \ell_t(\bar{\tau} \mid \{\tau_t\}) \right) \Big) \\
&\quad = \ln \mathbb{E}_{\tau} \exp\Big(- \eta_t \ell_t(\tau \mid \{\tau_t\}) + \eta_t \mathbb{E}_{\bar{\tau}} \ell_t(\bar{\tau} \mid \{\tau_t\}) \Big) \\
&\quad = \ln \mathbb{E}_{\tau} \exp\Big(- \eta_t \ell_t(\tau \mid \{\tau_t\}) \Big) + \ln \exp\Big(\eta_t \mathbb{E}_{\bar{\tau}} \ell_t(\bar{\tau} \mid \{\tau_t\}) \Big) \\
&\quad \leq \mathbb{E}_{\tau} \exp\left(-\eta_t \ell_t(\tau \mid \{\tau_t\}) \right) - 1 + \eta_t \mathbb{E}_{\bar{\tau}} \ell_t(\bar{\tau} \mid \{\tau_t\}) \qquad (28) \\
&\quad \leq \mathbb{E}_{\tau} \Big[\exp\left(-\eta_t \ell_t(\tau \mid \{\tau_t\}) \right) - 1 + \eta_t \ell_t(\tau \mid \{\tau_t\}) \Big] \\
&\quad \leq \mathbb{E}_{\tau} \eta_t^2 \frac{\ell_t(\tau \mid \{\tau_t\})^2}{2} \qquad (29) \\
&\quad = \frac{\eta_t^2}{2} \sum_{\tau \in \mathcal{H}} p_t(\tau) \ell_t(\tau \mid \{\tau_t\})^2 \\
&\quad = \frac{\eta_t^2}{2} \sum_{\tau \in \mathcal{H}} p_t(\tau) \left(\frac{\ell_t(\tau)}{p_t(\tau) + \gamma_t} \cdot \mathbb{1}(\tau_t = \tau) \right) \ell_t(\tau \mid \{\tau_t\}) \\
&\quad \leq \frac{\ell_{\max} \eta_t^2}{2} \sum_{\tau \in \mathcal{H}} \ell_t(\tau \mid \{\tau_t\}), \qquad (30)
\end{aligned}$$

where (28) holds due to $\ln(x) \leq x - 1$ for $x \leq 1$, and (29) uses $\exp(x) \leq 1 + x + x^2/2$ for $x \leq 0$.

Third step. Let $\tilde{L}_t(\tau) \coloneqq \sum_{i=1}^{t} \ell_i(\tau \mid \{\tau_i\})$ and $\tilde{L}_0(\tau) = 0$. Then, we have

$$
\begin{aligned}
&-\frac{1}{\eta_t} \ln \mathbb{E}_{\tau \sim p_t} \exp\Big(- \eta_t \ell_t(\tau \mid \{\tau_t\}) \Big) \\
&\qquad = -\frac{1}{\eta_t} \ln \sum_{\tau \in \mathcal{H}} p_t(\tau) \exp\Big(- \eta_t \ell_t(\tau \mid \{\tau_t\}) \Big) \\
&\qquad = -\frac{1}{\eta_t} \ln \sum_{\tau \in \mathcal{H}} \frac{\exp(-\eta_t \tilde{L}_{t-1}(\tau))}{\sum_{\bar{\tau} \in \mathcal{H}} \exp(-\eta_t \tilde{L}_{t-1}(\bar{\tau}))} \exp\Big(- \eta_t \ell_t(\tau \mid \{\tau_t\}) \Big) \\
&\qquad = -\frac{1}{\eta_t} \ln \frac{\sum_{\tau \in \mathcal{H}} \exp(-\eta_t \tilde{L}_t(\tau))}{\sum_{\tau \in \mathcal{H}} \exp(-\eta_t \tilde{L}_{t-1}(\tau))}. \qquad (31)
\end{aligned}
$$

Fourth step. We combine rewritten forms to continue the derivation as follows:

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{\tau \sim p_t} \ell(\tau \mid \{\tau_t\}) \\
&\qquad \le \sum_{t=1}^{T} \frac{\ell_{\max} \eta_t}{2} \sum_{\tau \in \mathcal{H}} \ell_t(\tau \mid \{\tau_t\}) - \sum_{t=1}^{T} \frac{1}{\eta_t} \ln \frac{\sum_{\tau \in \mathcal{H}} \exp\Big(-\eta_t \tilde{L}_t(\tau)\Big)}{\sum_{\tau \in \mathcal{H}} \exp\Big(-\eta_t \tilde{L}_{t-1}(\tau)\Big)} \qquad (32) \\
&\qquad \le \sum_{t=1}^{T} \frac{\ell_{\max} \eta_t}{2} \sum_{\tau \in \mathcal{H}} \ell_t(\tau \mid \{\tau_t\}) - \frac{1}{\eta_T} \sum_{t=1}^{T} \ln \frac{\sum_{\tau \in \mathcal{H}} \exp\Big(-\eta_t \tilde{L}_t(\tau)\Big)}{\sum_{\tau \in \mathcal{H}} \exp\Big(-\eta_t \tilde{L}_{t-1}(\tau)\Big)} \qquad (33) \\
&\qquad = \sum_{t=1}^{T} \frac{\ell_{\max} \eta_t}{2} \sum_{\tau \in \mathcal{H}} \ell_t(\tau \mid \{\tau_t\}) - \frac{1}{\eta_T} \ln \frac{\sum_{\tau \in \mathcal{H}} \exp\Big(-\eta_T \tilde{L}_T(\tau)\Big)}{\sum_{\tau \in \mathcal{H}} \exp\Big(-\eta_0 \tilde{L}_0(\tau)\Big)} \\
&\qquad = \sum_{t=1}^{T} \frac{\ell_{\max} \eta_t}{2} \sum_{\tau \in \mathcal{H}} \ell_t(\tau \mid \{\tau_t\}) - \frac{1}{\eta_T} \ln \sum_{\tau \in \mathcal{H}} \exp\Big(-\eta_T \tilde{L}_T(\tau)\Big) + \frac{\ln |\mathcal{H}|}{\eta_T} \\
&\qquad \le \sum_{t=1}^{T} \frac{\ell_{\max} \eta_T}{2} \sum_{\tau \in \mathcal{H}} \ell_t(\tau \mid \{\tau_t\}) - \frac{1}{\eta_t} \ln \Big(\max_{\tau} \exp\Big(-\eta_T \tilde{L}_T(\tau)\Big)\Big) + \frac{\ln |\mathcal{H}|}{\eta_T} \\
&\qquad = \sum_{t=1}^{T} \frac{\ell_{\max} \eta_t}{2} \sum_{\tau \in \mathcal{H}} \ell_t(\tau \mid \{\tau_t\}) + \min_{\tau} \sum_{t=1}^{T} \ell_t(\tau \mid \{\tau_t\}) + \frac{\ln |\mathcal{H}|}{\eta_T}, \qquad (34)
\end{aligned}
$$

where (32) holds from (27), (30), and (31). Also, (33) holds as η_t is non-increasing and $\frac{\sum_{\tau \in \mathcal{H}} \exp\left(-\eta_t \tilde{L}_t(\tau)\right)}{\sum_{\tau \in \mathcal{H}} \exp\left(-\eta_t \tilde{L}_{t-1}(\tau)\right)} \le 1$.

Note that $\mathbb{E}_{\tau \sim p_t} \ell(\tau \mid \{\tau_t\}) = \ell_t(\tau_t) - \gamma_t \sum_{\tau \in \mathcal{H}} \ell(\tau \mid \{\tau_t\})$ for the following reason:

$$
\begin{aligned}
\mathbb{E}_{\tau \sim p_t} \ell_t(\tau \mid \{\tau_t\}) &= \sum_{\tau \in \mathcal{H}} p_t(\tau) \frac{\ell_t(\tau)}{p_t(\tau) + \gamma_t} \mathbb{1}(\tau_t = \tau) \\
&= \sum_{\tau \in \mathcal{H}} \left(\frac{p_t(\tau) + \gamma_t}{p_t(\tau) + \gamma_t} \ell_t(\tau) \mathbb{1}(\tau_t = \tau) - \frac{\gamma_t}{p_t(\tau) + \gamma_t} \ell_t(\tau) \mathbb{1}(\tau_t = \tau) \right) \\
&= \ell_t(\tau_t) - \gamma_t \sum_{\tau \in \mathcal{H}} \ell_t(\tau \mid \{\tau_t\}).
\end{aligned}
$$

From this fact and (34), we have

$$\sum_{t=1}^{T}\ell_t(\tau_t)-\min_{\tau}\sum_{t=1}^{T}\ell_t(\tau\mid\{\tau_t\})\le\sum_{t=1}^{T}\frac{\ell_{\max}\eta_t}{2}\sum_{\tau\in\mathcal{H}}\ell_t(\tau\mid\{\tau_t\})+\sum_{t=1}^{T}\gamma_t\sum_{\tau\in\mathcal{H}}\ell(\tau\mid\{\tau_t\})+\frac{\ln|\mathcal{H}|}{\eta_T}$$
$$=\sum_{t=1}^{T}\left(\frac{\ell_{\max}\eta_t}{2}+\gamma_t\right)\sum_{\tau\in\mathcal{H}}\ell_t(\tau\mid\{\tau_t\})+\frac{\ln|\mathcal{H}|}{\eta_T}. \tag{35}$$

If $\eta_t\le\frac{2\gamma_t}{\ell_{\max}}$ then it satisfies the Lemma condition *i.e.,* $\alpha_{t,\tau}=\frac{\ell_{\max}\eta_t}{2}+\gamma_t\le 2\gamma_t$, we can apply Lemma 2, from (35) to get

$$\sum_{t=1}^{T}\ell_t(\tau_t)-\min_{\tau}\sum_{t=1}^{T}\ell_t(\tau\mid\{\tau_t\})\le\sum_{t=1}^{T}\left(\frac{\ell_{\max}\eta_t}{2}+\gamma_t\right)\sum_{\tau\in\mathcal{H}}\ell_t(\tau)+\frac{\ln|\mathcal{H}|}{\eta_T}+\ell_{\max}\ln\frac{2}{\delta}$$

with probability $1-\delta/2$.

After that, by Corollary 1, we get

$$\sum_{t=1}^{T}\ell_t(\tau_t)-\min_{\tau}\sum_{t=1}^{T}\ell_t(\tau)\le\sum_{t=1}^{T}\ell_t(\tau_t)-\min_{\tau}\sum_{t=1}^{T}\ell_t(\tau\mid\{\tau_t\})+\frac{\ell_{\max}}{2\gamma}\ln\frac{2|\mathcal{H}|}{\delta}$$

with probability $1-\delta/2$.

Finally, we get combining the above two via the union bound as follows:

$$\mathbf{Reg}_T=\sum_{t=1}^{T}\ell_t(\tau_t)-\min_{\tau}\sum_{t=1}^{T}\ell_t(\tau)\le\sum_{t=1}^{T}\left(\frac{\ell_{\max}\eta_t}{2}+\gamma_t\right)\sum_{\tau\in\mathcal{H}}\ell_t(\tau)+\frac{\ln|\mathcal{H}|}{\eta_T}+\ell_{\max}\ln\frac{2}{\delta}+\frac{\ell_{\max}}{2\gamma}\ln\frac{2|\mathcal{H}|}{\delta}$$

with probability $1-\delta$.

The concluding steps of our proof closely follow the analysis presented by [36]. Recall that if the learning rate $\eta_t\le\frac{2\gamma_t}{\ell_{\max}}$ it satisfies the conditions of Lemma 2. Then, letting $\eta_t=\frac{2\gamma_t}{\ell_{\max}}$ and $\gamma=\frac{\eta_T\ell_{\max}}{2}$ due to Corollary 1, we can bound the regret bound as follows:

$$\sum_{t=1}^{T}\left(\frac{\ell_{\max}\eta_t}{2}+\gamma_t\right)\sum_{\tau\in\mathcal{H}}\ell_t(\tau)+\frac{\ln|\mathcal{H}|}{\eta_T}+\ell_{\max}\ln\frac{2}{\delta}+\frac{\ell_{\max}}{2\gamma}\ln\frac{2|\mathcal{H}|}{\delta}$$
$$=\sum_{t=1}^{T}\ell_{\max}\eta_t\sum_{\tau\in\mathcal{H}}\ell_t(\tau)+\frac{2\ln|\mathcal{H}|}{\eta_T}+\ell_{\max}\ln\frac{2}{\delta}+\frac{1}{\eta_T}\ln\frac{2}{\delta}$$
$$\le\sum_{t=1}^{T}\ell_{\max}^2|\mathcal{H}|\eta_t+\frac{2\ln|\mathcal{H}|}{\eta_T}+\ell_{\max}\ln\frac{2}{\delta}+\frac{1}{\eta_T}\ln\frac{2}{\delta}. \tag{36}$$

First, we consider the setting with a known time horizon. In this case, we use a fixed learning rate, $\eta_t=\eta_T$ for all t, which simplifies the general regret bound to:

$$\mathbf{Reg}_T\le\ell_{\max}^2T|\mathcal{H}|\eta_T+\frac{2\ln|\mathcal{H}|}{\eta_T}+\ell_{\max}\ln\frac{2}{\delta}+\frac{1}{\eta_T}\ln\frac{2}{\delta}. \tag{37}$$

Finally, we can get the optimal learning rate $\eta_t=\sqrt{\frac{2\ln|\mathcal{H}|+\ln(2/\delta)}{\ell_{\max}^2T|\mathcal{H}|}}$. Here, to make the learning rate agnostic to δ, we approximate $\eta_t\approx\sqrt{\frac{2\ln|\mathcal{H}|}{\ell_{\max}^2T|\mathcal{H}|}}$, which follows the convention in the original analysis [36].

Then, from (37), we obtain the following upper bound:

$$\mathbf{Reg}_T\le\ell_{\max}\left(2\sqrt{2T|\mathcal{H}|\ln|\mathcal{H}|}+\left(1+\sqrt{\frac{T|\mathcal{H}|}{2\ln|\mathcal{H}|}}\right)\ln\frac{2}{\delta}\right)$$
$$=\mathcal{O}\left(2\ell_{\max}\sqrt{2T|\mathcal{H}|\ln|\mathcal{H}|/\delta}\right). \tag{38}$$

Second, for the setting with an unknown time horizon, we use the learning rate $\eta_t = \sqrt{\frac{\ln|\mathcal{H}|}{\ell_{\max}^2 t|\mathcal{H}|}}$. Then we get general regret bound from (36), noting that $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$

$$\begin{aligned}\mathbf{Reg}_T &\leq \ell_{\max}\left(4\sqrt{T|\mathcal{H}|\ln|\mathcal{H}|} + \left(1+\sqrt{\frac{T|\mathcal{H}|}{\ln|\mathcal{H}|}}\right)\ln\frac{2}{\delta}\right)\\ &= \mathcal{O}\left(4\ell_{\max}\sqrt{T|\mathcal{H}|\ln|\mathcal{H}|/\delta}\right).\end{aligned}$$

Expected regret The expected regret can be obtained via integrating the deviations in (38), *i.e.,*

$$\mathbb{E}[W] \leq \int_0^1 \frac{1}{\delta}\mathbb{P}\left(W > \ln\frac{1}{\delta}\right)d\delta.$$

Here, if we take $W = \frac{1}{\ell_{\max}\left(1+\sqrt{T|\mathcal{H}|/(2\ln|\mathcal{H}|)}\right)}\left(\mathbf{Reg}_T - 2\ell_{\max}\sqrt{2T|\mathcal{H}|\ln|\mathcal{H}|}\right) - \ln 2 > \ln\frac{1}{\delta}$, which holds with probability at most δ then

$$\mathbb{E}[W] \leq 1,$$

which suggests that

$$\begin{aligned}\mathbb{E}[\mathbf{Reg}_T] &\leq \ell_{\max}\left(2\sqrt{2T|\mathcal{H}|\ln|\mathcal{H}|} + (1+\ln 2)\left(1+\sqrt{\frac{T|\mathcal{H}|}{2\ln|\mathcal{H}|}}\right)\right)\\ &= \mathcal{O}(\ell_{\max}\sqrt{T|\mathcal{H}|\ln|\mathcal{H}|}).\end{aligned} \tag{39}$$

H.3 A Proof of Lemma 1

We derive a lower bound of $\mathbf{Reg}_T$ which allows us to bound the SR violation $\mathcal{V}_T$. From the definition of regret and the loss function $\ell_t(\tau,\alpha)$ in (4), the following inequality holds for any comparator policy $\tau \in \mathcal{H}$:

$$\begin{aligned}(1+\lambda)\mathbf{Reg}_T &= \sum_{t=1}^{T}\Big[a_t(\tau_t) + \lambda d_t(\tau_t,\alpha)\Big] - \min_{\tau'\in\mathcal{H}}\sum_{t=1}^{T}\left[a_t(\tau') + \lambda d_t(\tau',\alpha)\right]\\ &\geq \sum_{t=1}^{T}\Big[a_t(\tau_t) + \lambda d_t(\tau_t,\alpha)\Big] - \sum_{t=1}^{T}\left[a_t(\tau) + \lambda d_t(\tau,\alpha)\right].\end{aligned} \tag{40}$$

Let $\tau^* \in \mathcal{H}$ be the policy that minimizes the SR violation, *i.e.,* $\tau^* \in \arg\min_{\tau\in\mathcal{H}}\mathcal{V}_T(\tau)$, where $\mathcal{V}_T(\tau) \coloneqq \sum_{t=1}^{T} d_t(\tau,\alpha) - \alpha T$. By substituting $\tau = \tau^*$ into (40) and rearranging the terms regarding the constraint term d_t, we obtain:

$$\sum_{t=1}^{T}\lambda d_t(\tau_t,\alpha) - \sum_{t=1}^{T}\lambda d_t(\tau^*,\alpha) \leq (1+\lambda)\mathbf{Reg}_T + \sum_{t=1}^{T}\left[a_t(\tau^*) - a_t(\tau_t)\right]. \tag{41}$$

Using the definitions $\mathcal{V}_T(\tau) = \sum_{t=1}^{T} d_t(\tau,\alpha) - \alpha T$ and $\sum_{t=1}^{T} a_t(\tau_t) = T\cdot\mathbf{Ineff}_T$, we can rewrite (41) as:

$$\begin{aligned}\lambda\left(\mathcal{V}_T - \min_{\tau\in\mathcal{H}}\mathcal{V}_T(\tau)\right) &\leq (1+\lambda)\mathbf{Reg}_T + \sum_{t=1}^{T} a_t(\tau^*) - T\cdot\mathbf{Ineff}_T\\ &\leq (1+\lambda)\mathbf{Reg}_T + T(1-\mathbf{Ineff}_T),\end{aligned} \tag{42}$$

where the last inequality holds since $a_t(\cdot) \in [0,1]$, implying $\sum_{t=1}^{T} a_t(\tau^*) \leq T$.

Finally, we observe that the minimum SR violation is non-positive, as the fallback policy $\tau_\emptyset$ (always abstaining) achieves zero risk:

$$\min_{\tau\in\mathcal{H}}\mathcal{V}_T(\tau) \leq \mathcal{V}_T(\tau_\emptyset) = 0. \tag{43}$$

Dividing (42) by λ and applying this observation yields the final bound:

$$\mathcal{V}_T \leq \frac{(1+\lambda)\mathbf{Reg}_T}{\lambda} + \frac{T(1-\mathbf{Ineff}_T)}{\lambda}. \tag{44}$$

Setting $\lambda = T^{1/2}$, we conclude:

$$\mathcal{V}_T \leq \frac{(1+T^{1/2})\mathbf{Reg}_T}{T^{1/2}} + T^{1/2}(1-\mathbf{Ineff}_T) = \mathcal{O}(T^{1/2}). \tag{45}$$

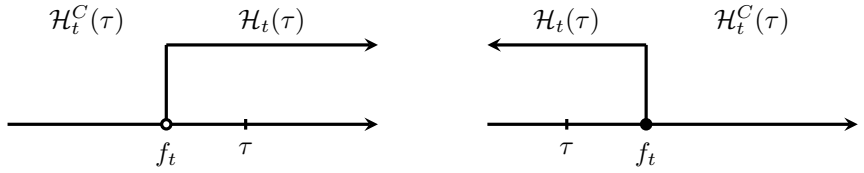
H.4 A Proof of Theorem 4

Here we prove Theorem 4 for an arbitrary loss range $\ell_t(\cdot) \in [0, \ell_{\max}]$ of the form (4) and a non-increasing learning rate; the theorem is the case $\ell_{\max} = 1$. Specifically, with $\eta_t = \frac{2\gamma_t}{\ell_{\max}} = \sqrt{\frac{\ln|\mathcal{H}|}{\ell_{\max}^2 T}}$, Algorithm 1 provides, with probability at least $1-\delta$,

$$\mathbf{Reg}_T \leq \ell_{\max}\left(4\sqrt{T\ln|\mathcal{H}|} + \left(1+\sqrt{\frac{T}{\ln|\mathcal{H}|}}\right)\ln\frac{2}{\delta}\right).$$

The proof technique follows that of the `Exp3-IX` [36] regret bound except that we use a novel loss estimator for feedback unlocking. Note that we abbreviate $\ell_t(\tau, \alpha \mid \mathcal{H}_t(\tau_t))$ as $\ell_t(\tau \mid \mathcal{H}_t(\tau_t))$ in this proof.

First, we introduce some properties of $\mathcal{H}_t(\tau_t)$ in our algorithm.



(a) Case for $\tau > f_t$ (b) Case for $\tau \leq f_t$

Figure 38: Visualization of $\mathcal{H}_t(\tau)$.

1. (self-inclusion) The first trivial property is that τ is in $\mathcal{H}_t(\tau)$ itself:
$$\tau \in \mathcal{H}_t(\tau). \tag{46}$$
2. (partition) The second property is that if $\tau' \in \mathcal{H}_t(\tau)$, then we have
$$\mathcal{H}_t(\tau') = \mathcal{H}_t(\tau). \tag{47}$$
This holds since if $\tau' \in \mathcal{H}_t(\tau)$ and $\tau \in \mathcal{H}_t(\tau)$ due to (46), τ and τ' are in the same side from f_t (*i.e.,* either $\tau \leq f_t \wedge \tau' \leq f_t$ or $\tau > f_t \wedge \tau' > f_t$), thus $\hat{S}(\mathbf{x}_t;\tau) = \hat{S}(\mathbf{x}_t;\tau')$, implying $\mathcal{H}_t(\tau') = \mathcal{H}_t(\tau)$ by our construction of $\mathcal{H}_t(\cdot)$ in our algorithm. The second case when $\tau \in \mathcal{H}_t(\tau')$ similarly holds as well.
3. (complement) The third property, similar to the second one, is that if $\tau' \in \mathcal{H}_t^C(\tau)$, then we have
$$\mathcal{H}_t^C(\tau) = \mathcal{H}_t(\tau'), \text{ or equivalently } \mathcal{H}_t(\tau) = \mathcal{H}_t^C(\tau'). \tag{48}$$
This property holds as follows: $\tau' \in \mathcal{H}_t(\tau')$ holds due to (46). Suppose $\tau' \in \mathcal{H}_t^C(\tau)$, then we have $\tau' \notin \mathcal{H}_t(\tau)$. This along with $\tau' \in \mathcal{H}_t(\tau')$ implies $\mathcal{H}_t(\tau') \neq \mathcal{H}_t(\tau)$. Since $\mathcal{H}_t(\tau_t)$ is either $\{\tau \mid \tau \leq f_t\}$ or $\{\tau \mid \tau > f_t\}$ depending on the value of τ_t, $\mathcal{H}_t(\cdot)$ partitions the space $[0,1]$ into two disjoint sets based on f_t, where one set is the complement of the other. Therefore, $\mathcal{H}_t(\tau) \neq \mathcal{H}_t(\tau')$ is equivalent to either $\mathcal{H}_t^C(\tau) = \mathcal{H}_t(\tau')$ or $\mathcal{H}_t(\tau) = \mathcal{H}_t^C(\tau')$.
4. (swap) The final property is that
$$\tau' \in \mathcal{H}_t(\tau) \quad \text{if and only if} \quad \tau \in \mathcal{H}_t(\tau'). \tag{49}$$
Suppose that $\tau' \in \mathcal{H}_t(\tau)$ and $\tau' \in \mathcal{H}_t(\tau')$, then the above holds as follows:
$$\tau \in \mathcal{H}_t(\tau) = \mathcal{H}_t(\tau'),$$
where the inclusion holds due to (46) and the equality holds due to (47) if $\tau', \tau \in \mathcal{H}_t(\tau)$. The reverse holds similarly as well.

With these properties, we can derive the following two useful equalities. Let

$$\hat{\ell}_t(\tau \mid \mathcal{H}_t(\tau_t)) = \frac{\ell_t(\tau)}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} \mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau})) \cdot p_t(\bar{\tau})} \mathbb{1}(\tau \in \mathcal{H}_t(\tau_t)), \tag{50}$$

which is our loss estimator without γ_t. Then, the following equality holds:

$$\begin{aligned}
\mathbb{E}_{\tau_t\sim p_t} \hat{\ell}_t(\tau \mid \mathcal{H}_t(\tau_t)) &= \sum_{\tau'\in\mathcal{H}} p_t(\tau') \left[\frac{\ell_t(\tau)}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau')} \mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau})) \cdot p_t(\bar{\tau})} \mathbb{1}(\tau \in \mathcal{H}_t(\tau')) \right] \\
&= \ell_t(\tau) \sum_{\tau'\in\mathcal{H}} \frac{p_t(\tau')}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau')} \mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau})) \cdot p_t(\bar{\tau})} \mathbb{1}(\tau \in \mathcal{H}_t(\tau')) \\
&= \ell_t(\tau) \sum_{\tau'\in\mathcal{H}} \frac{p_t(\tau')}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau')} \mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau})) \cdot p_t(\bar{\tau})} \mathbb{1}(\tau' \in \mathcal{H}_t(\tau)) && (51) \\
&= \ell_t(\tau) \sum_{\tau'\in\mathcal{H}_t(\tau)} \frac{p_t(\tau')}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau')} \mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau})) \cdot p_t(\bar{\tau})} && (52) \\
&= \ell_t(\tau) \sum_{\tau'\in\mathcal{H}_t(\tau)} \frac{p_t(\tau')}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau)} \mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau})) \cdot p_t(\bar{\tau})} && (53) \\
&= \ell_t(\tau) \sum_{\tau'\in\mathcal{H}_t(\tau)} \frac{p_t(\tau')}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau)} \mathbb{1}(\bar{\tau} \in \mathcal{H}_t(\tau)) \cdot p_t(\bar{\tau})} && (54) \\
&= \ell_t(\tau) \sum_{\tau'\in\mathcal{H}_t(\tau)} \frac{p_t(\tau')}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau)} p_t(\bar{\tau})} \\
&= \ell_t(\tau), && (55)
\end{aligned}$$

where (51) holds as (49), (52) follows from the properties of the summation and indicator function, (53) holds as (47) from $\tau' \in \mathcal{H}_t(\tau)$ in the summation, and (54) follows by the same argument as (51) from the property (49), *i.e.,* $\bar{\tau} \in \mathcal{H}_t(\tau)$ if $\tau \in \mathcal{H}_t(\bar{\tau})$.

Also, the following equality with respect to our loss estimator (8) holds similarly:

$$\begin{aligned}
\mathbb{E}_{\tau\sim p_t} \ell_t(\tau \mid \mathcal{H}_t(\tau_t)) &= \sum_{\tau\in\mathcal{H}} p_t(\tau) \frac{\ell_t(\tau)}{\gamma_t + \sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} \mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau})) p_t(\bar{\tau})} \cdot \mathbb{1}(\tau \in \mathcal{H}_t(\tau_t)) \\
&= \sum_{\tau\in\mathcal{H}_t(\tau_t)} p_t(\tau) \frac{\ell_t(\tau)}{\gamma_t + \sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} \mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau})) p_t(\bar{\tau})} \\
&= \ell_t(\tau_t) \sum_{\tau\in\mathcal{H}_t(\tau_t)} \frac{p_t(\tau)}{\gamma_t + \sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} \mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau})) p_t(\bar{\tau})} && (56) \\
&= \ell_t(\tau_t) \sum_{\tau\in\mathcal{H}_t(\tau_t)} \frac{p_t(\tau)}{\gamma_t + \sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} p_t(\bar{\tau})} && (57) \\
&= \ell_t(\tau_t) \frac{\sum_{\tau\in\mathcal{H}_t(\tau_t)} p_t(\tau)}{\gamma_t + \sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} p_t(\bar{\tau})} \\
&= \ell_t(\tau_t) \frac{\gamma_t + \sum_{\tau\in\mathcal{H}_t(\tau_t)} p_t(\tau)}{\gamma_t + \sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} p_t(\bar{\tau})} - \ell_t(\tau_t) \frac{\gamma_t}{\gamma_t + \sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} p_t(\bar{\tau})} && (58) \\
&= \ell_t(\tau_t) - \gamma_t \frac{\ell_t(\tau_t)}{\gamma_t + \sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} \mathbb{1}(\tau_t \in \mathcal{H}_t(\bar{\tau})) p_t(\bar{\tau})} \mathbb{1}(\tau_t \in \mathcal{H}_t(\tau_t)) && (59) \\
&= \ell_t(\tau_t) - \gamma_t \ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)), && (60)
\end{aligned}$$

where (56) holds as $\ell_t(\tau) = \ell_t(\tau')$ for all $\tau \in \mathcal{H}_t(\tau')$ by the definition of the specialized loss in (4), (57) holds as the same as (54), (58) holds as $\tau_t \in \mathcal{H}_t(\tau_t)$ and (49), (59) holds due to the property of the indicator function, and (60) holds by the definition.

Then, we first introduce a new Lemma 3 and Corollary 2 to establish a high probability bound.

Lemma 3. *Let $\alpha_t \le 2\gamma_t$ where γ_t is non-increasing in t. Then, the following inequality holds with probability at least $1-\delta$:*

$$\sum_{t=1}^{T} \alpha_t \ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)) - \sum_{t=1}^{T} 2\alpha_t \ell_{max} \le \ell_{max} \ln(1/\delta). \tag{61}$$

Proof. First, the following inequality holds for any $\tau \in \mathcal{H}$:

$$\begin{aligned}
\ell_t(\tau \mid \mathcal{H}_t(\tau_t)) &= \frac{\ell_t(\tau)}{\gamma_t + \sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} \mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau})) \cdot p_t(\bar{\tau})} \cdot \mathbb{1}(\tau \in \mathcal{H}_t(\tau_t)) \\
&= \ell_{\max} \cdot \frac{\ell_t(\tau)/\ell_{\max}}{\gamma_t + \sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} \mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau})) \cdot p_t(\bar{\tau})} \cdot \mathbb{1}(\tau \in \mathcal{H}_t(\tau_t)) \\
&\le \ell_{\max} \cdot \frac{\ell_t(\tau)/\ell_{\max}}{\gamma_t\ell_t(\tau)/\ell_{\max} + \sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} \mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau})) \cdot p_t(\bar{\tau})} \cdot \mathbb{1}(\tau \in \mathcal{H}_t(\tau_t)) \\
&= \frac{\ell_{\max}}{2\gamma_t} \cdot \frac{2\gamma_t\ell_t(\tau)/\ell_{\max}}{\gamma_t\ell_t(\tau)/\ell_{\max} + \sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} \mathbb{1}(\tau \in \mathcal{H}_t(\bar{\tau})) \cdot p_t(\bar{\tau})} \cdot \mathbb{1}(\tau \in \mathcal{H}_t(\tau_t)) \\
&= \frac{1}{\beta_t} \cdot \frac{\frac{\beta_t\ell_t(\tau)}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} \mathbb{1}(\tau\in\mathcal{H}_t(\bar{\tau}))\cdot p_t(\bar{\tau})}}{\frac{\beta_t\ell_t(\tau)}{2\sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} \mathbb{1}(\tau\in\mathcal{H}_t(\bar{\tau}))\cdot p_t(\bar{\tau})} + 1} \cdot \mathbb{1}(\tau \in \mathcal{H}_t(\tau_t)) \\
&= \frac{1}{\beta_t} \cdot \frac{\frac{\beta_t\ell_t(\tau)}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} \mathbb{1}(\tau\in\mathcal{H}_t(\bar{\tau}))\cdot p_t(\bar{\tau})} \cdot \mathbb{1}(\tau \in \mathcal{H}_t(\tau_t))}{\frac{\beta_t\ell_t(\tau)}{2\sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} \mathbb{1}(\tau\in\mathcal{H}_t(\bar{\tau}))\cdot p_t(\bar{\tau})} \cdot \mathbb{1}(\tau \in \mathcal{H}_t(\tau_t)) + 1} \\
&= \frac{1}{\beta_t} \cdot \frac{\beta_t\hat{\ell}_t(\tau \mid \mathcal{H}_t(\tau_t))}{\beta_t\hat{\ell}_t(\tau \mid \mathcal{H}_t(\tau_t))/2 + 1} \\
&\le \frac{1}{\beta_t} \cdot \ln\left(1 + \beta_t\hat{\ell}_t(\tau \mid \mathcal{H}_t(\tau_t))\right),
\end{aligned} \tag{62}$$

where $\beta_t = 2\gamma_t/\ell_{\max}$, $\hat{\ell}_t(\tau \mid \mathcal{H}_t(\tau_t)) \coloneqq \frac{\ell_t(\tau)}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} \mathbb{1}(\tau\in\mathcal{H}_t(\bar{\tau}))\cdot p_t(\bar{\tau})} \cdot \mathbb{1}(\tau \in \mathcal{H}_t(\tau_t))$, and the last inequality holds as $\frac{x}{1+x/2} \le \ln(1+x)$ for all $x \ge 0$.

Let $\mathbb{E}_t$ be the expectation conditioned on $\tau_1, \dots, \tau_{t-1}$ and τ' be any element in $\mathcal{H}$. Then, we have

$$
\begin{aligned}
&\mathbb{E}_t\left[\exp\left(\frac{\alpha_t}{\ell_{\max}}\ell_t(\tau_t \mid \mathcal{H}_t(\tau_t))\right)\right] \\
&\le \mathbb{E}_t\left[\exp\left(\frac{\alpha_t}{\ell_{\max}\beta_t}\ln\left(1+\beta_t\hat{\ell}_t(\tau_t \mid \mathcal{H}_t(\tau_t))\right)\right)\right] && (63)\\
&\le \mathbb{E}_t\left[1+\frac{\alpha_t}{\ell_{\max}}\hat{\ell}_t(\tau_t \mid \mathcal{H}_t(\tau_t))\right] && (64)\\
&= 1+\frac{\alpha_t}{\ell_{\max}}\mathbb{E}_t\hat{\ell}_t(\tau_t \mid \mathcal{H}_t(\tau_t)) \\
&= 1+\frac{\alpha_t}{\ell_{\max}}\sum_{\tau_t\in\mathcal{H}} p_t(\tau_t)\frac{\ell_t(\tau_t)}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)}\mathbb{1}(\tau_t\in\mathcal{H}_t(\bar{\tau}))\cdot p_t(\bar{\tau})}\cdot\mathbb{1}(\tau_t\in\mathcal{H}_t(\tau_t)) \\
&= 1+\frac{\alpha_t}{\ell_{\max}}\sum_{\tau_t\in\mathcal{H}} p_t(\tau_t)\frac{\ell_t(\tau_t)}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau_t)} p_t(\bar{\tau})} \\
&= 1+\frac{\alpha_t}{\ell_{\max}}\left(\sum_{\tau\in\mathcal{H}_t(\tau')}\frac{p_t(\tau)\ell_t(\tau)}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau)} p_t(\bar{\tau})}+\sum_{\tau\in\mathcal{H}_t^C(\tau')}\frac{p_t(\tau)\ell_t(\tau)}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau)} p_t(\bar{\tau})}\right) \\
&\le 1+\frac{\alpha_t}{\ell_{\max}}\ell_{\max}\left(\sum_{\tau\in\mathcal{H}_t(\tau')}\frac{p_t(\tau)}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau)} p_t(\bar{\tau})}+\sum_{\tau\in\mathcal{H}_t^C(\tau')}\frac{p_t(\tau)}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau)} p_t(\bar{\tau})}\right) \\
&= 1+\alpha_t\left(\sum_{\tau\in\mathcal{H}_t(\tau')}\frac{p_t(\tau)}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau')} p_t(\bar{\tau})}+\sum_{\tau\in\mathcal{H}_t^C(\tau')}\frac{p_t(\tau)}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau)} p_t(\bar{\tau})}\right) && (65)\\
&= 1+\alpha_t\left(\sum_{\tau\in\mathcal{H}_t(\tau')}\frac{p_t(\tau)}{\sum_{\bar{\tau}\in\mathcal{H}_t(\tau')} p_t(\bar{\tau})}+\sum_{\tau\in\mathcal{H}_t^C(\tau')}\frac{p_t(\tau)}{\sum_{\bar{\tau}\in\mathcal{H}_t^C(\tau')} p_t(\bar{\tau})}\right) && (66)\\
&= 1+2\alpha_t && (67)\\
&\le \exp(2\alpha_t), && (68)
\end{aligned}
$$

where (63 holds as (62). (64) uses $x\ln(1+y)\le\ln(1+xy)$ for all $y>-1$ and $x\in[0,1]$ since $\alpha_t\le 2\gamma_t=\ell_{\max}\beta_t$. (65) holds as (47), *i.e.,* $\mathcal{H}_t(\tau)=\mathcal{H}_t(\tau')$ due to $\tau\in\mathcal{H}_t(\tau')$ in the summation. Also, (66) holds due to (48), *i.e.,* $\mathcal{H}_t(\tau)=\mathcal{H}_t^C(\tau')$ as $\tau\in\mathcal{H}_t^C(\tau')$ from the summation condition. Finally, (67) holds because neither $\mathcal{H}_t(\tau)$ nor $\mathcal{H}_t^C(\tau)$ for $\tau\in\mathcal{H}$ can be empty under the definitions of f_t and $\mathcal{H}_t(\tau)$; see Figure 38.

Thus, we can find that $W_t \coloneqq \exp\left(\sum_{s=1}^t \frac{\alpha_s}{\ell_{\max}}\left(\ell_s(\tau_s \mid \mathcal{H}_s(\tau_s))-2\ell_{\max}\right)\right)$ is a supermartingale with respect to $\mathbb{E}_t$, *i.e.,* $\mathbb{E}_t[W_t]\le W_{t-1}$ as follows:

$$
\begin{aligned}
\mathbb{E}_t[W_t] &= \mathbb{E}_t\left[W_{t-1}\exp\left(\frac{\alpha_t}{\ell_{\max}}\left(\ell_t(\tau_t \mid \mathcal{H}_t(\tau_t))-2\ell_{\max}\right)\right)\right] \\
&= W_{t-1}\mathbb{E}_t\left[\exp\left(\frac{\alpha_t}{\ell_{\max}}\left(\ell_t(\tau_t \mid \mathcal{H}_t(\tau_t))-2\ell_{\max}\right)\right)\right] \\
&= W_{t-1}\mathbb{E}_t\left[\exp\left(\frac{\alpha_t}{\ell_{\max}}\ell_t(\tau_t \mid \mathcal{H}_t(\tau_t))\right)\right]\exp\left(-2\alpha_t\right) \\
&\le W_{t-1}\exp\left(2\alpha_t\right)\exp\left(-2\alpha_t\right) && (69)\\
&= W_{t-1},
\end{aligned}
$$

where (69) holds due to (68). Due to $W_0=1$ by the definition, the above implies $\mathbb{E}[W_T]\le 1$. Then, by applying this with the Markov's inequality, we have

$$
\mathbb{P}\left(\sum_{t=1}^T\frac{\alpha_t}{\ell_{\max}}\left(\ell_t(\tau_t \mid \mathcal{H}_t(\tau_t))-2\ell_{\max}\right)>\ln\frac{1}{\delta}\right)=\mathbb{P}\left(W_T>\frac{1}{\delta}\right)\le\delta\mathbb{E}_T[W_T]\le\delta.
$$

Thus,

$$\sum_{t=1}^{T} \frac{\alpha_t}{\ell_{\max}} \left(\ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)) - 2\ell_{\max}\right) \leq \ln(1/\delta),$$

with probability at least $1-\delta$, which completes the proof. □

Corollary 2. *Let $\gamma_t = \gamma > 0$ for all t. Then, simultaneously for all $\tau \in \mathcal{H}$,*

$$\sum_{t=1}^{T} \ell_t(\tau \mid \mathcal{H}_t(\tau_t)) - \sum_{t=1}^{T} \ell_t(\tau) \leq \frac{\ell_{max} \ln(|\mathcal{H}|/\delta)}{2\gamma}, \tag{70}$$

with probability at least $1-\delta$.

Proof. The proof of this corollary can be derived from Lemma 3, but here we provide a direct proof, which uses similar proof techniques as in Lemma 3.

Let $\gamma_t = \gamma \geq 0$. Then, we have

$$\begin{aligned}
\mathbb{E}_t\left[\exp\left(\frac{2\gamma}{\ell_{\max}}\ell_t(\tau \mid \mathcal{H}_t(\tau_t))\right)\right] &\leq \mathbb{E}_t\left[\exp\left(\frac{2\gamma}{\ell_{\max}\beta}\ln\left(1+\beta\hat{\ell}_t(\tau \mid \mathcal{H}_t(\tau_t))\right)\right)\right] && (71)\\
&\leq \mathbb{E}_t\left[1+\frac{2\gamma}{\ell_{\max}}\hat{\ell}_t(\tau \mid \mathcal{H}_t(\tau_t))\right] && (72)\\
&= 1+\frac{2\gamma}{\ell_{\max}}\ell_t(\tau) && (73)\\
&\leq \exp\left(\frac{2\gamma}{\ell_{\max}}\ell_t(\tau)\right), && (74)
\end{aligned}$$

where (71) holds same as (62) with fixed $\gamma_t = \gamma$ and $\beta_t = \beta$ for all t. Also, (72) uses $x\ln(1+y) \leq \ln(1+xy)$ for all $y > -1$ and $x \in [0,1]$ along with the fact that $2\gamma = \ell_{\max}\beta$ and thus $\frac{2\gamma}{\beta\ell_{\max}} = 1$. (73) holds as $\mathbb{E}_t\hat{\ell}_t(\tau \mid \mathcal{H}_t(\tau_t)) = \ell_t(\tau)$ from 55.

Thus, we can find that $W_t \coloneqq \exp\left(\sum_{s=1}^{t}\frac{2\gamma}{\ell_{\max}}\left(\ell_s(\tau \mid \mathcal{H}_s(\tau_s)) - \ell_s(\tau)\right)\right)$ is a supermartingale with respect to $\mathbb{E}_t$, *i.e.,* $\mathbb{E}[W_t] \leq W_{t-1}$ as follows:

$$\begin{aligned}
\mathbb{E}_t[W_t] &= \mathbb{E}_t\left[W_{t-1}\cdot\exp\left(\frac{2\gamma}{\ell_{\max}}\left(\ell_t(\tau \mid \mathcal{H}_t(\tau_t)) - \ell_t(\tau)\right)\right)\right]\\
&= W_{t-1}\mathbb{E}_t\left[\exp\left(\frac{2\gamma}{\ell_{\max}}\left(\ell_t(\tau \mid \mathcal{H}_t(\tau_t)) - \ell_t(\tau)\right)\right)\right]\\
&= W_{t-1}\mathbb{E}_t\left[\exp\left(\frac{2\gamma}{\ell_{\max}}\ell_t(\tau \mid \mathcal{H}_t(\tau_t))\right)\right]\exp\left(-\frac{2\gamma}{\ell_{\max}}\ell_t(\tau)\right)\\
&\leq W_{t-1}\exp\left(\frac{2\gamma}{\ell_{\max}}\ell_t(\tau)\right)\exp\left(-\frac{2\gamma}{\ell_{\max}}\ell_t(\tau)\right) && (75)\\
&= W_{t-1}.
\end{aligned}$$

where (75) holds as (74). From this and $W_0 = 1$ by the definition, we have $\mathbb{E}[W_T] \leq \mathbb{E}[W_{T-1}] \leq \cdots \leq 1$. By this and the Markov's inequality, we have

$$\mathbb{P}\left(\sum_{t=1}^{T}\frac{2\gamma}{\ell_{\max}}\left(\ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)) - \ell_t(\tau)\right) > \ln\frac{1}{\delta}\right) = \mathbb{P}\left(W_T > \frac{1}{\delta}\right) \leq \delta\mathbb{E}_T[W_T] \leq \delta.$$

Then, by the union bound, for all τ, we have

$$\sum_{t=1}^{T}\left(\ell_t(\tau \mid \mathcal{H}_t(\tau_t)) - \ell_t(\tau)\right) \leq \frac{\ell_{\max}\ln(|\mathcal{H}|/\delta)}{2\gamma}$$

with probability at least $1-\delta$, which completes the proof.

□

We then prove our main theorem, which consists of four steps.

First step. We split the expected loss estimator into two logarithmic terms Eq. (3.7) in [6]:

$$\mathbb{E}_{\tau\sim p_t}\ell_t(\tau \mid \mathcal{H}_t(\tau_t)) = \frac{1}{\eta_t}\underbrace{\ln \mathbb{E}_{\tau\sim p_t}\exp\Big(-\eta_t\Big(\ell_t(\tau \mid \mathcal{H}_t(\tau_t)) - \mathbb{E}_{\bar\tau\sim p_t}\ell_t(\bar\tau \mid \mathcal{H}_t(\tau_t))\Big)\Big)}_{\text{variability term}} - \frac{1}{\eta_t}\underbrace{\ln \mathbb{E}_{\tau\sim p_t}\exp\Big(-\eta_t\ell_t(\tau \mid \mathcal{H}_t(\tau_t))\Big)}_{\text{log of an exponential expectation term}}, \quad (76)$$

where η_t is non-increasing in $t \in \mathbb{N}$. In the following steps, each term will be rewritten.

Second step. We provide an upper bound on the first term:

$$\begin{aligned}
&\ln \mathbb{E}_{\tau\sim p_t}\exp\Big(-\eta_t\Big(\ell_t(\tau \mid \mathcal{H}_t(\tau_t)) - \mathbb{E}_{\bar\tau\sim p_t}\ell_t(\bar\tau \mid \mathcal{H}_t(\tau_t))\Big)\Big)\\
&\quad= \ln \mathbb{E}_{\tau\sim p_t}\exp\big(-\eta_t\ell_t(\tau \mid \mathcal{H}_t(\tau_t))\big) + \eta_t\mathbb{E}_{\bar\tau\sim p_t}\ell_t(\bar\tau \mid \mathcal{H}_t(\tau_t))\\
&\quad\le \mathbb{E}_{\tau\sim p_t}\left[\exp\big(-\eta_t\ell_t(\tau \mid \mathcal{H}_t(\tau_t))\big) - 1 + \eta_t\ell_t(\tau \mid \mathcal{H}_t(\tau_t))\right] &(77)\\
&\quad\le \mathbb{E}_{\tau\sim p_t}\eta_t^2\frac{\ell_t(\tau \mid \mathcal{H}_t(\tau_t))^2}{2} &(78)\\
&\quad\le \frac{\eta_t^2}{2}\sum_{\tau\in\mathcal{H}} p_t(\tau)\ell_t(\tau \mid \mathcal{H}_t(\tau_t))^2\\
&\quad= \frac{\eta_t^2}{2}\sum_{\tau\in\mathcal{H}} p_t(\tau)\left(\frac{\ell_t(\tau)}{\gamma_t + \sum_{\bar\tau\in\mathcal{H}_t(\tau_t)}\mathbb{1}(\tau\in\mathcal{H}_t(\bar\tau))\cdot p_t(\bar\tau)}\cdot\mathbb{1}(\tau\in\mathcal{H}_t(\tau_t))\right)^2 &(79)\\
&\quad= \frac{\eta_t^2}{2}\sum_{\tau\in\mathcal{H}_t(\tau_t)} p_t(\tau)\left(\frac{\ell_t(\tau)}{\gamma_t + \sum_{\bar\tau\in\mathcal{H}_t(\tau_t)}\mathbb{1}(\tau\in\mathcal{H}_t(\bar\tau))\cdot p_t(\bar\tau)}\right)^2\\
&\quad= \frac{\eta_t^2}{2}\sum_{\tau\in\mathcal{H}_t(\tau_t)} p_t(\tau)\left(\frac{\ell_t(\tau)}{\gamma_t + \sum_{\bar\tau\in\mathcal{H}_t(\tau_t)} p_t(\bar\tau)}\right)^2 &(80)\\
&\quad= \frac{\eta_t^2}{2}\frac{\sum_{\tau\in\mathcal{H}_t(\tau_t)} p_t(\tau)\ell_t(\tau)^2}{\left(\gamma_t + \sum_{\bar\tau\in\mathcal{H}_t(\tau_t)} p_t(\bar\tau)\right)^2}\\
&\quad\le \frac{\ell_{\max}\eta_t^2}{2}\frac{\sum_{\tau\in\mathcal{H}_t(\tau_t)} p_t(\tau)}{\left(\gamma_t + \sum_{\bar\tau\in\mathcal{H}_t(\tau_t)} p_t(\bar\tau)\right)^2}\ell_t(\tau_t) &(81)\\
&\quad\le \frac{\ell_{\max}\eta_t^2}{2}\frac{\ell_t(\tau_t)}{\gamma_t + \sum_{\bar\tau\in\mathcal{H}_t(\tau_t)} p_t(\bar\tau)}\\
&\quad= \frac{\ell_{\max}\eta_t^2}{2}\frac{\ell_t(\tau_t)}{\gamma_t + \sum_{\bar\tau\in\mathcal{H}_t(\tau_t)}\mathbb{1}(\tau_t\in\mathcal{H}_t(\bar\tau))\cdot p_t(\bar\tau)}\cdot\mathbb{1}(\tau_t\in\mathcal{H}_t(\tau_t)), &(82)\\
&\quad= \frac{\ell_{\max}\eta_t^2}{2}\ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)), &(83)
\end{aligned}$$

where (77) holds due to $\ln(x) \le x - 1$ for $x \le 1$, (78) uses $\exp(x) \le 1 + x + x^2/2$ for $x \le 0$, (80) holds as the same as (57), (81) holds as $\ell_t(\tau) = \ell_t(\tau')$ for all $\tau \in \mathcal{H}_t(\tau')$ by the definition of the specialized loss (4) and its domain $[0, \ell_{\max}]$, and (82) holds as the same as (59), *i.e.,* the property of the indicator function.

Third step. Next, we reformulate the second term, the log of an exponential expectation term. Let $\tilde{L}_t(\tau) \coloneqq \sum_{s=1}^{t} \ell_s(\tau \mid \mathcal{H}_s(\tau_s))$ and $\tilde{L}_0(\tau) = 0$. Then, we have

$$\begin{aligned}
&-\frac{1}{\eta_t} \ln \mathbb{E}_{\tau \sim p_t} \exp\Big(-\eta_t \ell_t(\tau \mid \mathcal{H}_t(\tau_t))\Big) \\
&\qquad = -\frac{1}{\eta_t} \ln \sum_{\tau \in \mathcal{H}} p_t(\tau) \exp\Big(-\eta_t \ell_t(\tau \mid \mathcal{H}_t(\tau_t))\Big) \\
&\qquad = -\frac{1}{\eta_t} \ln \sum_{\tau \in \mathcal{H}} \frac{\exp(-\eta_t \tilde{L}_{t-1}(\tau))}{\sum_{\bar{\tau} \in \mathcal{H}} \exp(-\eta_t \tilde{L}_{t-1}(\bar{\tau}))} \exp\Big(-\eta_t \ell_t(\tau \mid \mathcal{H}_t(\tau_t))\Big) \\
&\qquad = -\frac{1}{\eta_t} \ln \frac{\sum_{\tau \in \mathcal{H}} \exp(-\eta_t \tilde{L}_t(\tau))}{\sum_{\tau \in \mathcal{H}} \exp(-\eta_t \tilde{L}_{t-1}(\tau))}. \qquad (84)
\end{aligned}$$

Fourth step. Then, combining rewritten terms, (83) and (84), the following inequality holds:

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}_{\tau \sim p_t} \ell_t(\tau \mid \mathcal{H}_t(\tau_t)) &\le \sum_{t=1}^{T} \frac{\ell_{\max} \eta_t}{2} \ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)) - \sum_{t=1}^{T} \frac{1}{\eta_t} \ln \frac{\sum_{\tau \in \mathcal{H}} \exp(-\eta_t \tilde{L}_t(\tau))}{\sum_{\tau \in \mathcal{H}} \exp(-\eta_t \tilde{L}_{t-1}(\tau))} \\
&\le \sum_{t=1}^{T} \frac{\ell_{\max} \eta_t}{2} \ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)) - \frac{1}{\eta_T} \sum_{t=1}^{T} \ln \frac{\sum_{\tau \in \mathcal{H}} \exp(-\eta_t \tilde{L}_t(\tau))}{\sum_{\tau \in \mathcal{H}} \exp(-\eta_t \tilde{L}_{t-1}(\tau))} \qquad (85) \\
&= \sum_{t=1}^{T} \frac{\ell_{\max} \eta_t}{2} \ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)) - \frac{1}{\eta_T} \ln \frac{\sum_{\tau \in \mathcal{H}} \exp(-\eta_T \tilde{L}_T(\tau))}{\sum_{\tau \in \mathcal{H}} \exp(-\eta_1 \tilde{L}_0(\tau))} \\
&= \sum_{t=1}^{T} \frac{\ell_{\max} \eta_t}{2} \ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)) - \frac{1}{\eta_T} \ln \sum_{\tau \in \mathcal{H}} \exp(-\eta_T \tilde{L}_T(\tau)) + \frac{\ln |\mathcal{H}|}{\eta_T} \\
&\le \sum_{t=1}^{T} \frac{\ell_{\max} \eta_t}{2} \ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)) - \frac{1}{\eta_T} \ln \Big(\max_{\tau} \exp(-\eta_T \tilde{L}_T(\tau))\Big) + \frac{\ln |\mathcal{H}|}{\eta_T} \\
&= \sum_{t=1}^{T} \frac{\ell_{\max} \eta_t}{2} \ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)) + \min_{\tau} \sum_{t=1}^{T} \ell_t(\tau \mid \mathcal{H}_t(\tau_t)) + \frac{\ln |\mathcal{H}|}{\eta_T},
\end{aligned}$$

where (85) holds as η_t is non-increasing and $\ln \frac{\sum_{\tau \in \mathcal{H}} \exp(-\eta_t \tilde{L}_t(\tau))}{\sum_{\tau \in \mathcal{H}} \exp(-\eta_t \tilde{L}_{t-1}(\tau))} \le 0$ for all t.

From this and the fact (60) that $\mathbb{E}_{\tau \sim p_t} \ell_t(\tau \mid \mathcal{H}_t(\tau_t)) = \ell_t(\tau_t) - \gamma_t \ell_t(\tau_t \mid \mathcal{H}_t(\tau_t))$, we have

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(\tau_t) - \min_{\tau} \sum_{t=1}^{T} \ell_t(\tau \mid \mathcal{H}_t(\tau_t)) &\le \sum_{t=1}^{T} \frac{\ell_{\max} \eta_t}{2} \ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)) + \frac{\ln |\mathcal{H}|}{\eta_T} + \sum_{t=1}^{T} \gamma_t \ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)) \\
&= \sum_{t=1}^{T} \left(\frac{\ell_{\max} \eta_t}{2} + \gamma_t \right) \ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)) + \frac{\ln |\mathcal{H}|}{\eta_T}.
\end{aligned}$$

Then, from Lemma 3 and Corollary 2 along with the union bound, the following inequality holds with probability at least $1 - 2\delta'$, where $\delta' = \delta/2$ if $\eta_t \le \frac{2\gamma_t}{\ell_{\max}}$:

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(\tau_t) - \min_{\tau} \sum_{t=1}^{T} \ell_t(\tau) &\le \sum_{t=1}^{T} \left(\frac{\ell_{\max} \eta_t}{2} + \gamma_t \right) \ell_t(\tau_t \mid \mathcal{H}_t(\tau_t)) + \frac{\ln |\mathcal{H}|}{\eta_T} + \frac{\ell_{\max} \ln(2|\mathcal{H}|/\delta)}{2\gamma} \\
&\le 2\ell_{\max} \sum_{t=1}^{T} \left(\frac{\ell_{\max} \eta_t}{2} + \gamma_t \right) + \ell_{\max} \ln \frac{2}{\delta} + \frac{\ln |\mathcal{H}|}{\eta_T} + \frac{\ell_{\max} \ln(2|\mathcal{H}|/\delta)}{2\gamma},
\end{aligned}$$

where the first inequality holds with probability at least $1 - \delta'$ from Corollary 2 and the second inequality holds with probability at least $1 - \delta'$ from Lemma 3 when $\alpha_t = \frac{\ell_{\max} \eta_t}{2} + \gamma_t \le 2\gamma_t$.

Final Bounds. We have our regret bound depending on the setup of the learning rate η_t.

If $\eta_t = \frac{2\gamma_t}{\ell_{\max}} = \sqrt{\frac{\ln|\mathcal{H}|}{\ell_{\max}^2 T}}$, with probability at least $1-\delta$,

$$\begin{aligned}\mathbf{Reg}_T &\le 2\ell_{\max}\sqrt{T\ln|\mathcal{H}|} + \ell_{\max}\left(1+\sqrt{\frac{T}{\ln|\mathcal{H}|}}\right)\ln\frac{2}{\delta} + 2\ell_{\max}\sqrt{T\ln|\mathcal{H}|} \\ &= \ell_{\max}\left(4\sqrt{T\ln|\mathcal{H}|} + \left(1+\sqrt{\frac{T}{\ln|\mathcal{H}|}}\right)\ln\frac{2}{\delta}\right) \\ &= \mathcal{O}(\ell_{\max}\sqrt{T\ln(|\mathcal{H}|/\delta)}). \end{aligned} \tag{86}$$

If $\eta_t = \frac{2\gamma_t}{\ell_{\max}} = \sqrt{\frac{\ln|\mathcal{H}|}{2\ell_{\max}^2 t}}$, with probability at least $1-\delta$,

$$\begin{aligned}\mathbf{Reg}_T &\le 2\ell_{\max}\sqrt{2T\ln|\mathcal{H}|} + \ell_{\max}\left(1+\sqrt{\frac{2T}{\ln|\mathcal{H}|}}\right)\ln\frac{2}{\delta} + 2\ell_{\max}\sqrt{2T\ln|\mathcal{H}|} \\ &= \ell_{\max}\left(4\sqrt{2T\ln|\mathcal{H}|} + \left(1+\sqrt{\frac{2T}{\ln|\mathcal{H}|}}\right)\ln\frac{2}{\delta}\right) \\ &= \mathcal{O}(\ell_{\max}\sqrt{T\ln(|\mathcal{H}|/\delta)}), \end{aligned}$$

noting that $\sum_{t=1}^{T} 1/\sqrt{t} \le 2\sqrt{T}$.

Expected regret The expected regret can be obtained via integrating the deviations in (86), *i.e.,*

$$\mathbb{E}[W] \le \int_0^1 \frac{1}{\delta}\mathbb{P}\left(W > \ln\frac{1}{\delta}\right) d\delta.$$

Here, if we take $W = \frac{1}{\ell_{\max}\left(1+\sqrt{T/\ln|\mathcal{H}|}\right)}\left(\mathbf{Reg}_T - 4\ell_{\max}\sqrt{T\ln|\mathcal{H}|}\right) - \ln 2 > \ln\frac{1}{\delta}$, which holds with probability at most δ, then,

$$\mathbb{E}[W] \le 1,$$

which suggests that

$$\mathbb{E}[\mathbf{Reg}_T] \le \ell_{\max}\left(4\sqrt{T\ln|\mathcal{H}|} + (1+\ln 2)\left(1+\sqrt{\frac{T}{\ln|\mathcal{H}|}}\right)\right) = \mathcal{O}(\ell_{\max}\sqrt{T\ln|\mathcal{H}|}). \tag{87}$$

I Discussion

I.1 Inefficiency

The SR bound in (6) converges to zero in T unless $\mathbf{Ineff}_T$ converges to 1. In other words, $\mathbf{SR}_T$ may fail to be bounded if $\mathbf{Ineff}_T \approx 1$ *e.g.,* $\sum_{t=1}^{T} e_t \approx T$, due to the nature of the empirical metric with the variance induced by randomized bandit algorithms. For instance, if the generator G consistently produces incorrect outputs, the learner will try to converge to an always-abstaining solution. However, due to randomization, if the learner mistakenly accepts even a single sample in the early stage of learning, then regardless of convergence to an always-abstaining solution, the empirical SR may still exceed the desired level α. Figures 26(a) and 27(c) show such a case caused by a poorly calibrated scoring function rather than the generator: for NQ with GPT-5.4, $f_{\texttt{con}}$ assigns the maximum score to 44.5% of the samples, 14.8% of which are incorrect, so `ExAUL` is driven toward always abstaining for smaller α and the empirical SR exceeds α (see the caption of Figure 26).

I.2 Selecting λ

We clarify that given the data sequence $\mathbf{x}_{1:T}$, the realized selection inefficiency $\mathbf{Ineff}_T$ is an outcome of the interaction between the model capability of G, the calibration of the scoring function f, and the

choice of the trade-off parameter λ. In particular, a larger λ places more emphasis on the SR-related loss term (4), which typically leads to more conservative behavior and faster convergence of the SR constraint, at the cost of increased inefficiency.

Nevertheless, when the underlying model or the scoring function f is severely miscalibrated, one may heuristically increase λ so that the learner approaches the behavior of the most conservative SR-feasible expert in hindsight, if such an expert exists (see Section I.3 for a discussion on selection efficiency). In this sense, choosing $\lambda = \mathcal{O}(T^d)$ with $d \in \mathbb{R}^+$ provides a principled way to tune the trade-off between efficiency and robustness.

In our experiments, we set $\lambda = T^{1/2}$, matching the rate achieved by standard bandit regret bounds. Empirical trends under different values of λ are reported in Figures 30 and 31.

I.3 Selection Efficiency

In this section, we show that the always-abstaining policy is not optimal in the aspect of full-feedback setting whenever there exists a selectively-abstaining policy that satisfies the SR constraint, *i.e.,* $\arg\min_\tau \sum_{t=1}^T [a_t(\tau) + \lambda d_t(\tau, \alpha)] \notin \mathcal{H}^{\text{aa}}$, if there exists some $\tau \in \mathcal{H} \setminus \mathcal{H}^{\text{aa}}$ that satisfies the SR guarantee, where $\mathcal{H}^{\text{aa}} \coloneqq \{\tau \mid \sum_{t=1}^T a_t(\tau) = T\}$.

Let $e_t^\tau \coloneqq \mathcal{E}_t(\hat{S}(;\tau))$, and $\tau^{\text{aa}} \in \mathcal{H}^{\text{aa}}$, which is always abstaining, then for any $\tau^{\text{sa}} \in \mathcal{H} \setminus \mathcal{H}^{\text{aa}}$ that satisfies the SR guarantee, the following inequality holds,

$$\frac{\sum_{t=1}^T \mathbb{1}\left(\hat{S}(\mathbf{x}_t; \tau^{\text{sa}}) \neq \texttt{IDK}\right) \cdot e_t^{\tau^{\text{sa}}}}{\sum_{t=1}^T \mathbb{1}\left(\hat{S}(\mathbf{x}_t; \tau^{\text{sa}}) \neq \texttt{IDK}\right)} \leq \alpha$$

$$\Rightarrow \sum_{t=1}^T \mathbb{1}\left(\hat{S}(\mathbf{x}_t; \tau^{\text{sa}}) \neq \texttt{IDK}\right) \cdot e_t^{\tau^{\text{sa}}} \leq \alpha \sum_{t=1}^T \mathbb{1}\left(\hat{S}(\mathbf{x}_t; \tau^{\text{sa}}) \neq \texttt{IDK}\right). \tag{88}$$

Let $L_T(\tau, \alpha) \coloneqq \sum_{t=1}^T \frac{a_t(\tau) + \lambda d_t(\tau, \alpha)}{1+\lambda}$, which is the cumulative loss of τ, and consider the difference of cumulative loss between τ^{aa} and τ^{sa},

$$\begin{aligned}
&(1+\lambda)(L_T(\tau^{\text{aa}}) - L_T(\tau^{\text{sa}})) \\
&= \sum_{t=1}^T [a_t(\tau^{\text{aa}}) + \lambda d_t(\tau^{\text{aa}}, \alpha)] - \sum_{t=1}^T [a_t(\tau^{\text{sa}}) + \lambda d_t(\tau^{\text{sa}}, \alpha)] \\
&= \sum_{t=1}^T \left[\mathbb{1}\left(\hat{S}(\mathbf{x}_t; \tau^{\text{aa}}) = \texttt{IDK}\right) + \lambda \left(\mathbb{1}(\hat{S}(\mathbf{x}_t; \tau^{\text{aa}}) \neq \texttt{IDK}) \cdot e_t^{\tau^{\text{aa}}} - \alpha \mathbb{1}(\hat{S}(\mathbf{x}_t; \tau^{\text{aa}}) \neq \texttt{IDK}) + \alpha \right) \right] \\
&\quad - \sum_{t=1}^T \left[\mathbb{1}\left(\hat{S}(\mathbf{x}_t; \tau^{\text{sa}}) = \texttt{IDK}\right) + \lambda \left(\mathbb{1}(\hat{S}(\mathbf{x}_t; \tau^{\text{sa}}) \neq \texttt{IDK}) \cdot e_t^{\tau^{\text{sa}}} - \alpha \mathbb{1}(\hat{S}(\mathbf{x}_t; \tau^{\text{sa}}) \neq \texttt{IDK}) + \alpha \right) \right] \\
&= T + \lambda\alpha T - \left[T - \sum_{t=1}^T \mathbb{1}\left(\hat{S}(\mathbf{x}_t; \tau^{\text{sa}}) \neq \texttt{IDK}\right) + \lambda \left\{ \alpha \left(T - \sum_{t=1}^T \mathbb{1}\left(\hat{S}(\mathbf{x}_t; \tau^{\text{sa}}) \neq \texttt{IDK}\right) \right) \right.\right. \\
&\qquad \left.\left. + \sum_{t=1}^T \mathbb{1}\left(\hat{S}(\mathbf{x}_t; \tau^{\text{sa}}) \neq \texttt{IDK}\right) \cdot e_t^{\tau^{\text{sa}}} \right\} \right] \\
&= \sum_{t=1}^T \mathbb{1}\left(\hat{S}(\mathbf{x}_t; \tau^{\text{sa}}) \neq \texttt{IDK}\right) + \lambda\alpha \sum_{t=1}^T \mathbb{1}\left(\hat{S}(\mathbf{x}_t; \tau^{\text{sa}}) \neq \texttt{IDK}\right) - \lambda \sum_{t=1}^T \mathbb{1}\left(\hat{S}(\mathbf{x}_t; \tau^{\text{sa}}) \neq \texttt{IDK}\right) \cdot e_t^{\tau^{\text{sa}}}.
\end{aligned}$$

Since τ^{sa} satisfies the SR guarantee (88), we get

$$L_T(\tau^{\text{aa}}) - L_T(\tau^{\text{sa}}) \geq \sum_{t=1}^T \frac{\mathbb{1}\left(\hat{S}(\mathbf{x}_t; \tau^{\text{sa}}) \neq \texttt{IDK}\right)}{1+\lambda} = \frac{T - \sum_{t=1}^T a_t(\tau^{\text{sa}})}{1+\lambda} > 0, \tag{89}$$

where $T - \sum_{t=1}^T a_t(\tau^{\text{sa}}) > 0$ holds due to $\tau^{\text{sa}} \notin \mathcal{H}^{\text{aa}}$. This implies that, given sufficiently many data points, any τ^{sa} that satisfies the SR guarantee incurs at least $T - \sum_{t=1}^T a_t(\tau^{\text{sa}})$ less loss than τ^{aa}. This is empirically supported by Figure 26 and 27.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims and contributions are stated in the abstract and Section 1.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations of the proposed approach in Section 5.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification:

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Sections 4 and Appendix E provide the experimental setup, implementation details, hyperparameters, datasets, baselines, and evaluation protocols needed to reproduce the main results.

   Guidelines:

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

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

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

Answer: [No]

Justification: We do not provide open access to the code and data at submission time. However, Sections 4 and Appendix E provide the implementation details, hyperparameters, and evaluation protocols needed to reproduce the main results.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: Sections 4 and Appendix E specify the experimental setup, including data splits, hyperparameters, implementation details, and evaluation protocols needed to understand the results.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: We report distributional information for the main experimental results using violin plots in Section 4, which show variability across repeated runs. We also provide theoretical guarantees in Section 3.4 to support the reliability of the proposed algorithm.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We describe the compute resources used for our experiments in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research was conducted in accordance with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: The work is foundational algorithmic research and does not have a direct path to societal impact or deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release data or models with a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We properly cite the original dataset sources and respect their licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets such as datasets, code packages, or models.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs were not used as important, original, or non-standard components of the core method.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.