COFACT: CONFORMAL FACTUALITY GUARANTEES FOR LANGUAGE MODELS UNDER COVARIATE SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) excel in natural language processing (NLP) tasks but often generate false or misleading information, known as hallucinations, raising reliability concerns in high-stakes applications. To provide statistical guarantees on the factuality of LLM outputs, conformal prediction based techniques have been proposed. Despite their strong theoretical guarantees, they rely heavily on the exchangeability assumption between calibration and test data, which is frequently violated in real-world scenarios with dynamic covariate shifts. To overcome this limitation, we introduce **CoFact**, a conformal prediction framework that uses online density ratio estimation to adaptively reweigh calibration data, ensuring alignment with evolving test distributions. With this approach, CoFact bypasses the exchangeability requirement and provides robust factuality guarantees under non-stationary conditions. To theoretically justify CoFact, we establish an upper bound on the gap between the actual hallucination rate and the target level α, demonstrating that the bound asymptotically approaches zero as the number of rounds and calibration samples increase. Empirically, CoFact is evaluated on MedLFQA, WikiData, and the newly introduced **WildChat+** dataset, which captures real-world covariate shifts through user-generated prompts. Results demonstrate that CoFact consistently outperforms existing methods, maintaining reliability even under dynamic conditions.

1 INTRODUCTION

Large Language Models (LLMs) have demonstrated exceptional performance across a wide range of natural language processing (NLP) tasks (Touvron et al., 2023; Devlin et al., 2019; OpenAI, 2023). Despite their impressive capabilities, their reliability and trustworthiness remain significant concerns. A critical issue is hallucination, where LLMs generate false or misleading information (Nadeau et al., 2024) that can lead to severe consequences in sensitive areas like healthcare (Jung, 2025), finance (Kang & Liu, 2023), and legal advice (Dahl et al., 2024). These limitations pose significant barriers to the broader adoption of LLMs in critical applications.

To address this issue, recent research have sought to improve the reliability of their outputs (Lewis et al., 2020; Chuang et al., 2023; Nakano et al., 2022). While these methods enhance factuality, they fall short of providing precise statistical guarantees, which are essential for high-stakes applications. To bridge this gap, a line of work has explored the use of conformal prediction to establish statistical guarantees on the factuality of LLM outputs. Specifically, Mohri & Hashimoto (2024) proposed splitting LLM-generated outputs into atomic sub-claims and filtering out those with factuality scores below a threshold determined via conformal inference, thereby offering marginal guarantees on factuality. Building on this, Cherian et al. (2024) extended the framework to provide subgroup-specific guarantees using conditional conformal prediction (Gibbs et al., 2025).

Although conformal inference-based methods provide factuality guarantees, they rely heavily on the assumption of exchangeability between calibration and test data (Vovk et al., 2005). In practice, however, this assumption is frequently violated due to factors such as topic drift (Reimer et al., 2023) and changes in user composition (Li et al., 2023). For example, Reimer et al. (2023) observed that the frequency of Covid-19-related terms in user queries fluctuated dramatically during the pandemic, exhibiting distinct peaks and troughs over time. Similar dynamics are also found in our analysis of real-world user prompt data, where topics evolve rapidly, as discussed in Section 5.2. Under such conditions, the distribution of test prompts can deviate significantly from that of the calibration set, thereby violating the exchangeability assumption.

Since the exchangeability assumption underpins the theoretical reliability of conformal prediction, its violation fundamentally compromises these guarantees, rendering such methods ineffective for ensuring reliable factuality in dynamic, real-world settings. To address this critical limitation, in this paper, we aim to answer the following question:

Given a stream of prompts from an unknown and dynamically evolving distribution, how can we provide factuality guarantees for the outputs of LLMs?

Answering this question presents two key challenges: (1) test samples arrive sequentially under continuous covariate shifts, making conformal prediction methods handling covariate shift that require static density ratio estimation across the entire test dataset infeasible (Tibshirani et al., 2019); and (2) in our scenario, unlike existing online conformal prediction methods (Gibbs & Candes, 2021; Gibbs & Candès, 2024; Areces et al., 2025), we do not have access to ground-truth labels for test samples after predictions, which prevents direct application of these methods.

To address these challenges, we introduce **CoFact**, a novel conformal prediction framework that integrates techniques from online learning into conformal prediction framework seamlessly. CoFact handles the continuous shifting by employing an online density ratio estimation mechanism that dynamically activates and updates multiple expert models across different time intervals. This adaptive approach enables CoFact to effectively track and learn the evolving density ratios between calibration and test distributions in real time. Leveraging these density ratio estimates, CoFact strategically reweighs calibration examples to align with the shifting test distribution. Through such integration, CoFact bypasses the traditional exchangeability assumption, providing robust reliability guarantees for LLMs even under a continual shifting prompt stream.

We theoretically establish an upper bound on the gap between the actual hallucination rate and the user-specified hallucination level α under shifting distributions, demonstrating that this bound asymptotically approaches zero as the number of rounds and calibration samples increase. This analysis offers a novel perspective that addresses limitations in existing methodologies.

To empirically demonstrate the effectiveness of CoFact, we evaluate CoFact on two well-established benchmarks, MedLFQA and WikiData, as well as a newly introduced benchmark, **WildChat+**. Built upon WildChat (Zhao et al., 2023), WildChat+ includes prompts generated by real users, effectively capturing real-world covariate shifts and enabling a more comprehensive evaluation. Experimental results demonstrate that CoFact significantly outperforms existing conformal prediction methods that rely on the exchangeability assumption, consistently maintaining factuality guarantees even under dynamically shifting distributions. These findings underscore CoFact's effectiveness in providing reliability guarantees in complex and dynamic real-world scenarios.

In summary, our contributions are as follows:

- **Novel Framework:** We propose **CoFact**, a conformal prediction framework designed to provide reliability guarantees for LLMs in the presence of continually shifting distributions.
- **Theoretical Guarantees:** We present rigorous theoretical analysis, establishing an upper bound on the gap between the actual hallucination rate and the target level α under shifting distributions. This analysis provides a solid foundation for CoFact's reliability in dynamic environments.
- **New Dataset:** To enable robust evaluation in real-world scenarios, we introduce **WildChat+**, which contains real user prompts along with LLM-generated responses and factuality annotations.
- **Extensive Experiments:** We conduct a comprehensive set of experiments across multiple benchmarks, including MedLFQA, WikiData, and WildChat+. The results demonstrate CoFact's effectiveness in maintaining reliability guarantees under diverse and evolving distributions.

2 Preliminaries

Conformal Prediction Conformal prediction is a statistical framework that converts the outputs of a black-box predictor (e.g., a single label) into prediction sets (e.g., a list of labels) that are guaranteed to include the true label with a user-specified confidence level of $1 - \alpha$. Formally, given an i.i.d. calibration set $\{(X_i, Y_i)\}_{i=1}^n$, where X_i and Y_i represent the features and labels, respectively, and a test sample X_{n+1} that is exchangeable with the calibration data, conformal prediction constructs a prediction set $\hat{C}(X_{n+1})$ such that the true label Y_{n+1} is included with probability at least $1-\alpha$:

$$\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1-\alpha. \tag{1}$$

To introduce the basic ideas behind conformal inference, we first define some notation, following Tibshirani et al. (2019). We denote by Quantile$(\alpha; \Psi)$ the level-α quantile of a distribution Ψ. Formally, for $X \sim \Psi$, the quantile is defined as:

$$\text{Quantile}(\alpha; \Psi) = \inf\{x : \mathbb{P}[X \leq x] \geq \alpha\}.$$

For empirical distributions, we denote the quantile of a multiset of values $v_1, \ldots, v_n$ as:

$$\text{Quantile}(\alpha; \{v_i\}_{i=1}^n) = \text{Quantile}\left(\alpha; \frac{1}{n}\sum_{i=1}^{n} \delta_{v_i}\right),$$

where δ_a represents a point mass at a (i.e., the distribution that places all its probability mass at a).

The central idea of conformal prediction involves the use of conformity score $V(X_i, Y_i)$, which quantifies how well the label Y_i corresponds to the features X_i. Using this score, the prediction set $\hat{C}(X_{n+1})$ is constructed by including all candidate labels y for which the conformity score $V(X_{n+1}, y)$ exceeds or meets a threshold τ. Given the conformity scores from the calibration set $\{V(X_i, Y_i)\}_{i=1}^n$, the threshold τ is determined as the $(1-\alpha)$-quantile of these scores combined with $\{\infty\}$ to ensure proper coverage. Formally,

$$\tau = \text{Quantile}\left(1-\alpha; \{V(X_i, Y_i)\}_{i=1}^n \cup \{\infty\}\right). \tag{2}$$

Conformal Factuality Control To generate a single response with guaranteed factuality, rather than a prediction set containing multiple potential factual responses, Mohri & Hashimoto (2024) proposes treating each response as a set of atomic claims and using conformal prediction to filter out hallucinated claims. Specifically, their method assumes access to an annotated calibration set consisting of n i.i.d. prompt-response-claim-label tuples, denoted as $D_0 = \{(P_i, R_i, \mathbf{C}_i, \mathbf{W}_i)\}_{i=1}^n$. Here, P_i represents the prompt for sample i, R_i is the corresponding response generated by the LLM, $\mathbf{C}_i = \{C_{i,j}\}_{j=1}^{k_i}$ denotes the set of claims extracted from R_i, and $\mathbf{W}_i = \{W_{i,j}\}_{j=1}^{k_i}$ represents the binary factuality labels for each claim, where $W_{i,j} = 1$ indicates that $C_{i,j}$ is factual, and $W_{i,j} = 0$ indicates that it is hallucinated.

The objective is to output a filtered response $F(\mathbf{C}_{n+1})$ for a test sample (P_{n+1}, R_{n+1}), which is exchangeable with the calibration data, such that the probability of including any hallucinated claims is bounded by a pre-defined level α:

$$\mathbb{P}\left(\exists C_{n+1,j} \in F(\mathbf{C}_{n+1}) \text{ such that } W_{n+1,j} = 0\right) \leq \alpha. \tag{3}$$

The filtered response $F(\mathbf{C}_{n+1})$ is constructed by excluding claims with low factuality scores:

$$F(\mathbf{C}_{n+1}) = \{C_{n+1,j} \in \mathbf{C}_{n+1} \mid p(C_{n+1,j}, P_{n+1}) \geq \tau\}, \tag{4}$$

where τ is the $(1-\alpha)$-quantile of the conformity scores $\{V(\mathbf{C}_i, \mathbf{W}_i)\}_{i=1}^n \cup \{\infty\}$. The conformity score $V(\mathbf{C}_i, \mathbf{W}_i)$ is defined as:

$$V(\mathbf{C}_i, \mathbf{W}_i) = \inf\left\{\tau \mid \forall C_{i,j} \in F(\mathbf{C}_i), W_{i,j} = 1\right\}, \tag{5}$$

and $p(C_{n+1,j}, P_{n+1})$ represents the factuality score, which measures how likely the claim $C_{n+1,j}$ is to be factual given the prompt P_{n+1}.

Building on the above framework, Cherian et al. (2024) argue that the guarantee provided by conformal factuality control is only marginal, meaning it applies globally across all test samples but does not account for specific subgroups of data. To address this limitation, they propose a group-wise guarantee inspired by conditional conformal prediction (Gibbs et al., 2025). This approach ensures that the factuality guarantee holds for all subgroups $G \in \mathcal{G}$, where the groups are defined by a family of functions. Specifically, the group-wise guarantee ensures:

$$\mathbb{P}\left(\exists C_{n+1,j} \in F(\mathbf{C}_{n+1}) \text{ such that } W_{n+1,j} = 0 \mid Z_{n+1} \in G\right) \leq \alpha \quad \text{for all } G \in \mathcal{G}. \tag{6}$$

Table 1: Glossary of commonly used symbols.

Symbol	Meaning	Symbol	Meaning
P_i	i-th prompt	Z_i	Tuple $(P_i, R_i, \mathbf{C}_i)$
R_i	Response to P_i generated by the LLM	D_0, $\mathcal{D}_0$	Calibration dataset/distribution
$\mathbf{C}_i$	Claims parsed from R_i	D_t, $\mathcal{D}_t$	Test dataset/distribution at time t
$\mathbf{W}_i$	Factuality labels of $\mathbf{C}_i$	r_t^*, $\hat{r}_t$	True/estimated density ratio
$C_{i,j}$	j-th claim of $\mathbf{C}_i$	w_t^*, $\hat{w}_t$	True/estimated importance weights
$W_{i,j}$	Factuality label of $C_{i,j}$	F, $\hat{F}$	Filtered sub-claims using $w_t^*/\hat{w}_t$

3 Methodology

In this section, we provide a detailed introduction to our proposed framework, CoFact. We begin by outlining the problem setup, including the continual covariate shift settings and our goal. Next, we present an oracle method that assumes access to the true density ratio to address the covariate shift. Lastly, we introduce our practical algorithm, which leverages online density ratio estimation to operate effectively in real-world scenarios.

3.1 Problem Setup

For clarity and notational simplicity, we define Z_i as the prompt-response pair, i.e., $Z_i = (P_i, R_i, \mathbf{C}_i)$. Thus, each sample in the calibration set can be represented as $(Z_i, \mathbf{W}_i)$. We consider an online setting with covariate shift, where we initially have access to a calibration set D_0 of size n, independently drawn from an initial distribution $\mathcal{D}_0$. At each subsequent round $t \in [T] \triangleq \{1, \ldots, T\}$, an unlabeled dataset D_t of size n_t is independently sampled from the current distribution $\mathcal{D}_t$, which may evolve continuously over time. For simplicity, and without loss of generality, we assume $n_t = 1$, representing the test sample arriving at time t as Z_{n+t}. To address the challenges posed by the covariate shift, we introduce the following assumption:

Assumption 1 (Continuous Covariate Shift)**.** *For any $Z \in \mathcal{Z}$ in the prompt-response space and any round $t \in [T]$, the conditional distribution of $\mathbf{W}$ given Z remains unchanged, i.e.,*

$$\mathcal{D}_t(\mathbf{W} \mid Z) = \mathcal{D}_0(\mathbf{W} \mid Z),$$

and the density ratio between $\mathcal{D}_t$ and $\mathcal{D}_0$ satisfies:

$$r_t^*(Z) = \frac{\mathcal{D}_t(Z)}{\mathcal{D}_0(Z)} \leq B < \infty.$$

Objective Our objective is to generate a filtered response $\hat{F}(\mathbf{C}_{n+t})$ for each test sample Z_{n+t} at round t, ensuring that the probability of including any hallucinated claims remains below a predefined threshold α. Given the challenges of providing exact guarantees at each time step under non-stationary distributions, we adopt the metric of prior works (Gibbs & Candes, 2021; Gibbs & Candès, 2024) and focus on bounding the gap between the average hallucination rate over T rounds and the target level α: $\left| \frac{1}{T} \sum_{t=1}^{T} \widehat{\mathrm{err}}_t - \alpha \right|$, where the error indicator for round t is defined as:

$$\widehat{\mathrm{err}}_t = \mathbb{1}\left[\exists C_{n+t,i} \in \hat{F}(\mathbf{C}_{n+t}) \text{ such that } W_{n+t,i} = 0\right]. \tag{7}$$

Here, $\hat{F}$ denotes the filtered response constructed by our method, distinguishing it from F, which is constructed using the true density ratio. The latter will be introduced in the next subsection. To aid understanding, Table 1 provides a glossary of commonly used symbols.

3.2 Conformal Factuality Control Under Covariate Shift with Oracle

We first consider the ideal scenario where the true density ratio $r_t^*(Z)$ is available for all $t \in [T]$. In this case, a standard approach to address covariate shift is to reweigh the calibration samples and the test sample using the density ratio when calculating the threshold τ_t (Tibshirani et al., 2019). Formally, given the conformity scores computed on the calibration set $\{V_i\}_{i=1}^n = \{V(Z_i, \mathbf{W}_i)\}_{i=1}^n$ and the test sample Z_{n+t}, the threshold τ_t at any round $t \in [T]$ is defined as:

$$\tau_t = \text{Quantile}(1 - \alpha; \sum_{i=1}^{n} w_t^*(Z_i)\delta_{V_i} + w_t^*(Z_{n+t})\delta_{\infty}), \tag{8}$$

where w_t^* is the weight function derived from the normalized density ratio:

$$w_t^*(Z) = \frac{r_t^*(Z)}{\sum_{i=1}^n r_t^*(Z_i) + r_t^*(Z_{n+t})}. \tag{9}$$

Using this threshold, the filtered response is constructed as:

$$F(\mathbf{C}_{n+t}) = \{C_{n+t,j} \in \mathbf{C}_{n+t} \mid p(C_{n+t,j}, P_{n+t}) \geq \tau_t\}, \tag{10}$$

where $p(C_{n+t,j}, P_{n+t})$ represents the factuality score of the j-th claim given the prompt P_{n+t}.

Corollary 1. *Given a calibration set D_0 and a test sample Z_{n+t} independent with D_0, if the true density ratio $r_t^*(Z)$ is available for all $t \in [T]$, then the filtered response constructed using Equation 10 with the threshold defined by Equation 8 satisfies the following guarantee:*

$$\mathbb{P}\left(\exists C_{n+t,i} \in F(\mathbf{C}_{n+t}) \textit{ such that } W_{n+t,i} = 0\right) \leq \alpha. \tag{11}$$

This result directly follows from Theorem 1 in Mohri & Hashimoto (2024) and Corollary 1 in Tibshirani et al. (2019).

3.3 Conformal Factuality Control with Online DRE

While guarantees on the hallucination rate can be established under the assumption of access to the true density ratios, the true density ratios are typically inaccessible in practice—particularly in scenarios where the underlying distribution is continuously evolving. To address this challenge, we adapt the method proposed in Zhang et al. (2023) to estimate a sequence of density ratios, $\{\hat{r}_t\}_{t=1}^T$, that approximate the true density ratios, $\{r_t^*\}_{t=1}^T$, under a dynamically changing distribution. In this section, we first reformulate the problem of online density ratio estimation (DRE) as a dynamic regret minimization problem. Next, we provide a brief overview of the online ensemble method employed to minimize dynamic regret. Finally, we describe how the estimated density ratios are integrated into the CoFact framework.

3.3.1 Online DRE via Dynamic Regret Minimization

As shown by Sugiyama et al. (2012), the problem of density ratio estimation can be reformulated as a Bregman divergence minimization problem. Consequently, to accurately estimate the density ratio at each time step $t \in [T]$, we solve the following optimization problem to obtain $\hat{r}_t$:

$$\min_{r \in \mathcal{H}_r} L_t^\psi(r) - L_t^\psi(r_t^*), \tag{12}$$

where L_t^ψ is the loss function for the density ratio, defined as:

$$L_t^\psi(r) = \mathbb{E}_{Z\sim\mathcal{D}_0}\left[\partial\psi(r(Z))r(Z) - \psi(r(Z))\right] - \mathbb{E}_{Z\sim\mathcal{D}_t}\left[\partial\psi(r(Z))\right]. \tag{13}$$

Here, ψ is the associated divergence function. By choosing different forms of ψ, various existing density ratio estimation methods can be recovered, including LSIF (Kanamori et al., 2009), the logistic regression method (Bickel et al., 2009), and UKL (Nguyen et al., 2007).

Building on this single-round density ratio estimation, it is natural to construct a sequence of estimators $\{\hat{r}_t\}_{t\in[T]}$ that perform well over time by minimizing the cumulative loss gap:

$$\sum_{t=1}^T \left(L_t^\psi(\hat{r}_t) - L_t^\psi(r_t^*)\right).$$

Implementation To implement this optimization, we make the following design choices:

- **Function Class and Divergence Function Specification**: We instantiate the density ratio function class $\mathcal{H}_r$ as a logistic regression model:

  $$\mathcal{H}_r \triangleq \mathcal{H}_\theta = \left\{\mathbf{z} \mapsto \exp(-\phi(\mathbf{z})^\top\theta) \;\middle|\; \|\phi(\mathbf{z})\|_2 \leq R, \|\theta\|_2 \leq S\right\},$$

  i.e., we model the density ratio estimator $\hat{r}_t$ as $\hat{r}_t(\cdot) = \exp(-\phi(\cdot)^\top\hat{\theta}_t)$, where $\phi(\mathbf{z})$ is a feature mapping function (e.g., the representation extracted by a neural network), and $\hat{\theta}_t$ is the parameter corresponding to $\hat{r}_t$. The bounded norms of $\phi(\mathbf{z})$ and θ ensure that the generalization gap can be analyzed. Moreover, we choose the divergence function ψ as:

  $$\psi = \psi_{\text{LR}} \triangleq t\log t - (t+1)\log(t+1).$$

- **Empirical Risk Minimization**: Since the true distributions $\mathcal{D}_0$ and $\mathcal{D}_t$ are inaccessible in practice, we use samples from a calibration set $D_0 = \{Z_i\}_{i=1}^n$ and a test set D_t. At each time step $t \in [T]$, $\hat{r}_t$ is obtained by solving the following empirical risk minimization problem:

$$\min_{\theta \in \Theta} \sum_{t=1}^{T} \hat{L}_t(\theta) - \hat{L}_t(\theta_t^*), \tag{14}$$

where Θ denotes the parameter space, θ_t^* is the optimal parameter corresponding to the true density ratio r_t^*, defined as $r_t^*(\cdot) = \exp(-\phi(\cdot)^\top \theta_t^*)$, and

$$\hat{L}_t(\theta) = \mathbb{E}_{Z\sim D_0}\left[\partial\psi(r(Z;\theta))r(Z;\theta) - \psi(r(Z;\theta))\right] - \mathbb{E}_{Z\sim D_t}\left[\partial\psi(r(Z;\theta))\right]. \tag{15}$$

Based on the above design choices, we are actually finding a sequence of parameters $\{\hat{\theta}_t\}_{t=1}^T$ to minimize the empirical dynamic regret in Equation 14.

3.3.2 Online Ensemble Framework for Dynamic Regret Minimization

To find the parameter sequence that minimizes dynamic regret, we adopt the online ensemble framework proposed by Zhang et al. (2023), which maintains a pool of experts. Each expert estimates the density ratio over its designated lifetime, and predictions from all active experts are aggregated to construct a global model at each time step, providing the final density ratio estimation. The framework operates through three key steps at each time step:

1. **Active-set update**: Experts are initialized with lifetimes chosen geometrically $(2^0, 2^1, 2^2, \ldots)$, and are re-initialized upon the expiration of their lifetimes.
2. **Model aggregation**: The parameters of active experts are weighted based on their historical performance and aggregated to form the global model $\hat{\theta}_t$. This aggregation step enables the global model to adaptively emphasize different segments of historical data, thereby enhancing its ability to capture covariate shifts.
3. **Expert update**: Active experts update their parameters $\theta_{t,i}$ using an online Newton step (ONS) method, which minimizes the regret $\hat{L}_t^\psi(\theta_{t,i}) - \hat{L}_t^\psi(\theta_t^*)$ at the current time step.

For a comprehensive description of the algorithm, please refer to Appendix C.

3.3.3 The Overall Framework

After obtaining the density ratio estimator $\hat{r}_t$ parameterized by $\hat{\theta}$ at time step t, we substitute it for the true density ratio r_t^* in Equation 8 to compute the threshold $\hat{\tau}_t$. This threshold is then used to filter hallucinated claims in the response:

$$\hat{\tau}_t = \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n} \hat{w}_t(Z_i)\delta_{V_i} + \hat{w}_t(Z_{n+t})\delta_\infty\right), \tag{16}$$

where $\hat{w}_t(Z)$ is the normalized estimated density ratio:

$$\hat{w}_t(Z) = \frac{\hat{r}_t(Z)}{\sum_{i=1}^{n} \hat{r}_t(Z_i) + \hat{r}_t(Z_{n+t})}. \tag{17}$$

Filtered responses are then given by:

$$\hat{F}(\mathbf{C}_{n+t}) = \{C_{(n+t)j} \in \mathbf{C}_{n+t} \mid p(C_{(n+t)j}, P_{n+t}) \geq \hat{\tau}_t\}. \tag{18}$$

4 Theoretical Guarantee

To obtain the theoretical guarantee on the hallucination rate, we need to make the following assumptions on the function class of the density ratio estimator r_t^* and the property of the divergence function ψ.

Assumption 2. *The true density ratio r_t^* is contained in the hypothesis space as $r_t^* \in \mathcal{H}_r = \mathcal{H}_\theta^{LR} \triangleq \{\mathbf{z} \mapsto \exp(-\phi(\mathbf{z})^\top\theta) \mid \theta \in \Theta\}$ for any $t \in [T]$ and the norm of θ and $\phi(\mathbf{z})$ are bounded by S and R respectively, i.e., $\|\theta\|_2 \leq S$ and $\|\phi(\mathbf{z})\|_2 \leq R$.*

Assumption 3. *The divergence function ψ is μ-strongly convex function satisfying $t\partial^3\psi(t) \leq 0$ and $\partial^3\psi(t) \leq 0$ for all $t \in$ dom ψ.*

This assumption can be satisfied by many commonly used divergence functions such as $\psi_{\text{LS}}(t) = (t-1)^2/2$ and $\psi_{\text{LR}}(t) = t\log t - \log(t+1)$ when the input is bounded, which is guaranteed by Assumption 2.

Theorem 1. *Under the assumptions 1, 2 and 3, with probability at least $1-\delta$, the gap between the averaged hallucination rate over T time steps and the target level α is bounded as*

$$\left| \frac{1}{T}\sum_{t=1}^{T} \widehat{\text{err}}_t - \alpha \right| \leq \tilde{\mathcal{O}}\left(\max\left\{ T^{-\frac{2}{3}} V_T^{\frac{2}{3}}, T^{-\frac{1}{2}} \right\} + 1/n \right) \tag{19}$$

when the parameter of the online ensemble is properly set. Here, $V_T = \sum_{t=2}^{T} \|\mathcal{D}_t(\mathbf{z}) - \mathcal{D}_{t-1}(\mathbf{z})\|_1$ measures the variation of input densities and the notation $\tilde{\mathcal{O}}$ hides logarithmic factors of T and $1/\delta$.

In this theorem, we can observe that the gap converges to 0 as the time horizon T and calibration set size n increase, and the convergence rate depends on the variation of input densities. This observation is consistent with our intuition that the more drastic the covariate shift is, the harder it is to adapt to the changing distribution. The proof of Theorem 1 is provided in Appendix D.2.

5 Experiments

In this section, we demonstrate the effectiveness of CoFact through experiments on both simulated and real-world covariate shifts. For all experiments, we set the target factuality level to $1-\alpha = 0.9$. We compare CoFact against the following baseline methods: (1) **SCP** (Mohri & Hashimoto, 2024), which employs standard conformal prediction to provide marginal factuality guarantees; and (2) **CondCP** (Cherian et al., 2024), which uses conditional conformal prediction to achieve group-wise factuality guarantees. To assess the performance of each method, we use two key metrics: **Factuality** and **Claims Retained**, defined as follows:

$$\textbf{Factuality} = 1 - \frac{1}{T}\sum_{t=1}^{T} \widehat{\text{err}}_t, \quad \text{and} \quad \textbf{Claims Retained} = \frac{1}{T}\sum_{t=1}^{T} \frac{|\hat{F}(\mathbf{C}_{n+t})|}{|\mathbf{C}_{n+t}|}. \tag{20}$$

5.1 Results on Simulated Covariate Shifts

Datasets We evaluate our method under simulated continual covariate shifts using two established datasets: **MedLFQA** (Jeong et al., 2024) and **WikiData** (Cherian et al., 2024). The MedLFQA dataset is a long-form medical question-answering benchmark with answers given by experts or LLMs, which are used to evaluate the factuality for sub-claims. WikiData is constructed by generating short biographies for sampled Wikipedia names. The factuality of sub-claims is evaluated through an adapted FAcTscore procedure, leveraging evidence from Wikipedia passages.

Since neither MedLFQA nor WikiData naturally exhibits covariate shifts, we simulate such shifts as follows. The dataset is first randomly divided into a calibration set (D_0) and a test set (D_{test}) of the same size. Then, at each time step t, the test samples Z_{n+t} are drawn from D_{test} according to a time-varying distribution $\mathcal{D}_t$, which is defined as a mixture of two base distributions, $\mathcal{D}'$ and $\mathcal{D}''$. To emulate continual covariate shifts, we define four patterns for $\mathcal{D}_t$: periodic shifts following sine (**Sin**) or square wave (**Squ**) patterns, gradual linear transitions from $\mathcal{D}'$ to $\mathcal{D}''$ over T time steps (**Lin**), and rapid stochastic alternations between $\mathcal{D}'$ and $\mathcal{D}''$ based on a fixed probability (**Ber**). Additional details on the dataset construction and shift simulation procedures are provided in Appendix E.1 and Appendix E.2.

Results We conduct experiments on MedLFQA and WikiData under the four types of simulated covariate shifts for $T = 2000$ time steps. The results for MedLFQA and WikiData are summarized in Table 2 and Table 3. Several key observations can be drawn from these tables. First, SCP experiences a significant drop in factuality under all types of covariate shifts across both datasets and fails to achieve the target factuality level of 0.9. This highlights the vulnerability of SCP when the exchangeability assumption is violated, which can severely degrade its performance. Second, while CondCP achieves high factuality on the MedLFQA dataset, it suffers from an extremely low claims retention rate. Additionally, CondCP exhibits very low factuality on the WikiData dataset,

Table 2: Averaged factuality and claims retained on the MedLFQA dataset under four types of shifts. Values in the range [0.89, 0.91] are highlighted in bold. Each experiment is repeated five times with different random seeds, and the results are reported as the mean ± standard deviation.

	Lin		Squ		Sin		Ber	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
SCP	0.811 ± 0.014	0.912 ± 0.011	0.808 ± 0.017	0.911 ± 0.012	0.810 ± 0.014	0.910 ± 0.010	0.828 ± 0.022	0.910 ± 0.006
CondCP	0.940 ± 0.004	0.389 ± 0.028	0.937 ± 0.007	0.400 ± 0.030	0.939 ± 0.005	0.394 ± 0.030	0.949 ± 0.010	0.364 ± 0.057
CoFact	**0.895 ± 0.026**	0.715 ± 0.031	**0.897 ± 0.022**	0.718 ± 0.030	**0.894 ± 0.018**	0.715 ± 0.031	**0.900 ± 0.019**	0.714 ± 0.036

Table 3: Averaged factuality and claims retained on WikiData. Settings are the same as Table 2.

	Lin		Squ		Sin		Ber	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
SCP	0.884 ± 0.006	0.780 ± 0.005	0.883 ± 0.006	0.781 ± 0.005	0.883 ± 0.006	0.781 ± 0.004	0.875 ± 0.010	0.782 ± 0.005
CondCP	0.724 ± 0.010	0.910 ± 0.002	0.726 ± 0.009	0.910 ± 0.003	0.725 ± 0.010	0.910 ± 0.002	0.716 ± 0.007	0.909 ± 0.004
CoFact	**0.896 ± 0.010**	0.748 ± 0.006	**0.895 ± 0.009**	0.748 ± 0.006	**0.895 ± 0.008**	0.748 ± 0.006	**0.897 ± 0.008**	0.749 ± 0.006

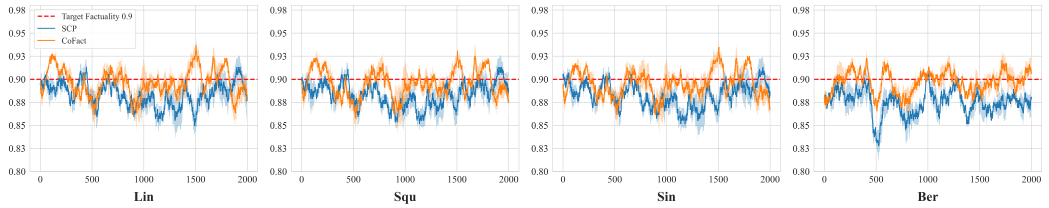


Figure 1: Factuality over time on the WikiData dataset. Each subplot corresponds to a different type of covariate shift, with the X-axis denoting the time steps and the Y-axis representing factuality. Factuality is computed using a sliding window that includes 50 steps before and after each time step. The curve shows the mean across 5 runs, while the shaded area indicates the standard deviation.

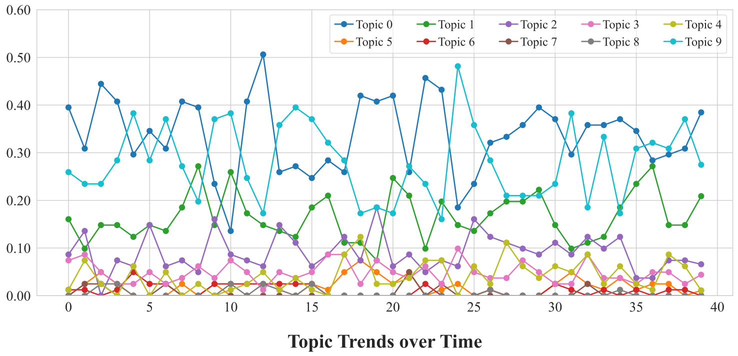


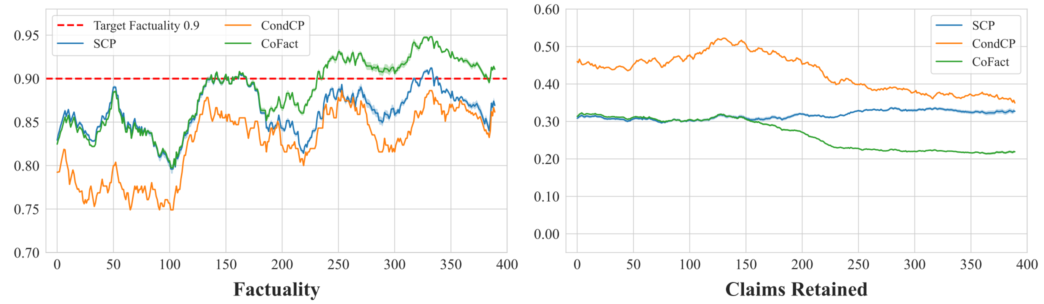
Figure 2: Topic proportions over time on WildChat+. X-axis represent the index of time intervals and Y-axis represent the proportion of each topic. Each line represent a topic identified by BERTopic.

indicating that group-wise factuality guarantees alone are insufficient for maintaining robust performance in shifting environments. Finally, among the three methods evaluated, our proposed method consistently achieves factuality closest to the target level of 0.9 under all types of covariate shifts across both datasets, demonstrating its effectiveness in adapting to dynamic distribution changes.

Furthermore, we visualize how factuality evolves over time on the WikiData dataset in Figure 1. Due to CondCP's significantly lower factuality compared to SCP and our method, we exclude its results from the figure for clarity. To produce smooth curves, factuality is calculated using a sliding window that spans 50 steps before and after the current time step. The figure reveals that our method consistently maintains factuality near the target level of 0.9 over time. Notably, the curve representing our method remains above that of SCP, particularly beyond time step 1000, further underscoring the advantage of our approach in adapting to shifting distributions.

5.2 Results on Real-world Covariate Shifts

Dataset and Analysis To evaluate our method in a real-world shifting setting, we construct a new benchmark **WildChat+** from WildChat Zhao et al. (2023), which contains user-generated prompts in the wild. For further construction details of the dataset, please refer to Appendix E.1. We conduct



Factuality

Claims Retained

Figure 3: Factuality and retained claims ratio over time on the WildChat+ dataset. The X-axis represents the time steps, while the Y-axis displays the factuality and retained claims ratio. Factuality is computed in the same manner as described in Figure 1. The curves represent the mean across 5 runs, and the shaded areas indicate the standard deviation.

data analysis on the WildChat+ dataset to show that the topics of prompts in the dataset change over time. Specifically, we first use the Latent Dirichlet Allocation (LDA) algorithm to identify 10 topics in the dataset and then split the dataset into 40 time intervals according to the timestamps of the conversations. After that, we plot the proportion of each topic in each time interval in Figure 2. From the figure, we can observe that the topics in the dataset change over time, which demonstrates the existence of a continual covariate shift in the dataset.

Results To evaluate our method under real-world covariate shifts, we perform experiments on the WildChat+ dataset. We split the data into two parts according to their timestamp: the first 40% of the data is used as the calibration set, while the remaining 60% serves as the test set. The results are presented in Figure 3. From the figure, several observations can be made. First, both SCP and CondCP struggle to reach the target factuality level of 0.9, underscoring the need for methods designed for dynamic conditions. Second, while CoFactinitially performs similarly to SCP before time step 200, it progressively adapts to the changing distribution and achieves factuality near the target level of 0.9 after time step 200. This demonstrates the effectiveness of our method in handling evolving distributions. It is worth noting that although CondCP achieves higher claims retention compared to SCP and our method, its overall factuality remains significantly lower. As a result, CondCP fails to meet the primary objective of the task: ensuring a reliable factuality guarantee.

Case Study To demonstrate the effectiveness of our method, we present a concrete example based on the filtered response to the prompt: ”What is Visual Studio Code?” The filtered claims is expressed by red strikethrough text.

Visual Studio Code is a free, ~~open-source~~ code editor developed by Microsoft. It is a ~~lightweight yet~~ powerful tool that supports various programming languages and offers features such as syntax highlighting, code completion, ~~debugging,~~ and Git integration. Visual Studio Code is highly customizable through extensions ~~and themes~~, making it popular among developers for writing ~~and debugging code~~.

From this filtered response, we can see that our method successfully removes the hallucinated claim ”open-source” while preserving the majority of the accurate information. This example highlights the capability of our approach to mitigate hallucinations in LLM-generated responses. Due to space constraints, we provide another case study in Appendix G.

6 Conclusion

In this paper, we tackle the critical challenge of providing factuality guarantees for LLMs in the presence of dynamic, real-world covariate shifts. To address the limitations of existing methods that rely on the exchangeability assumption, we introduce **CoFact**, a novel conformal prediction framework that utilizes online density ratio estimation to adaptively reweigh calibration data, ensuring alignment with evolving test distributions. Through both theoretical analysis and empirical evaluation, we demonstrate that CoFact consistently outperforms existing approaches in maintaining reliable factuality guarantees under dynamic and non-stationary conditions. The discussion of the limitations and future work can be found in Appendix H.

7 Ethics Statement

This paper introduces WildChat+, a derived dataset based on WildChat, which consists of real-world user-generated prompts. Due to the nature of real-world data, the dataset may contain personal information or potentially harmful content. While WildChat employs measures such as anonymization and the removal of sensitive information to address these concerns, it is still possible that some such content remains. Consequently, WildChat+ may also include similar issues. We strongly encourage users to handle the dataset responsibly and exercise caution. Beyond the concerns outlined above, we do not foresee any additional ethical issues associated with this study.

8 Reproducibility Statement

We have made significant efforts to ensure the reproducibility of our results. The code required to reproduce the experiments presented in this paper is included in the supplementary materials, and the implementation details are thoroughly described in Appendix E.3. Additionally, the detailed processing and construction procedures for our dataset are thoroughly described in the Appendix E.1. All assumptions underlying our theoretical results are clearly stated in Section 3.1 and 4 of the main text, and complete proofs of these results are provided in the Appendix D.2.

References

Amit Agarwal, Elad Hazan, Satyen Kale, and Robert E. Schapire. Algorithms for portfolio management based on the Newton method. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 9–16, New York, NY, USA, June 2006. Association for Computing Machinery. ISBN 978-1-59593-383-6. doi: 10.1145/1143844.1143846.

Felipe Areces, Christopher Mohri, Tatsunori Hashimoto, and John Duchi. Online Conformal Prediction via Online Optimization. In *Forty-Second International Conference on Machine Learning*, June 2025.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved Online Conformal Prediction via Strongly Adaptive Online Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 2337–2363. PMLR, July 2023.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative Learning Under Covariate Shift. *Journal of Machine Learning Research*, 10(75):2137–2155, 2009. ISSN 1533-7928.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. Conformal Prediction for Natural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516, 2024. doi: 10.1162/tacl_a_00715.

John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842, December 2024.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. In *The Twelfth International Conference on Learning Representations*, October 2023.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, 16(1):64–93, January 2024. ISSN 2161-7201. doi: 10.1093/jla/laae003.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 6491–6501, New York, NY, USA, August 2024. Association for Computing Machinery. ISBN 979-8-4007-0490-1. doi: 10.1145/3637528.3671470.

Isaac Gibbs and Emmanuel Candes. Adaptive Conformal Inference Under Distribution Shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672. Curran Associates, Inc., 2021.

Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. qkaf008, March 2025. ISSN 1369-7412. doi: 10.1093/jrsssb/qkaf008.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, August 2007. ISSN 1573-0565. doi: 10.1007/s10994-007-5016-8.

Minbyul Jeong, Hyeon Hwang, Chanwoong Yoon, Taewhoo Lee, and Jaewoo Kang. OLAPH: Improving Factuality in Biomedical Long-form Question Answering, October 2024.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active Retrieval Augmented Generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495.

Kyu-Hwan Jung. Large Language Models in Medicine: Clinical Applications, Technical Challenges, and Ethical Considerations. *Healthcare Informatics Research*, 31(2):114–124, April 2025. ISSN 2093-3681. doi: 10.4258/hir.2025.31.2.114.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A Least-squares Approach to Direct Importance Estimation. *Journal of Machine Learning Research*, 10(48):1391–1445, 2009. ISSN 1533-7928.

Haoqiang Kang and Xiao-Yang Liu. Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination, November 2023.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal Prediction with Large Language Models for Multi-Choice Question Answering, July 2023.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N. Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality Enhanced Language Models for Open-Ended Text Generation. *Advances in Neural Information Processing Systems*, 35:34586–34599, December 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020.

Moxin Li, Wenjie Wang, Fuli Feng, Yixin Cao, Jizhi Zhang, and Tat-Seng Chua. Robust Prompt Optimization for Large Language Models Against Distribution Shifts. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1539–1554, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.95.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741.

Christopher Mohri and Tatsunori Hashimoto. Language Models with Conformal Factuality Guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 36029–36047. PMLR, July 2024.

David Nadeau, Mike Kroutikov, Karen McNeil, and Simon Baribeau. Benchmarking Llama2, Mistral, Gemma and GPT for Factuality, Toxicity, Bias and Propensity for Hallucinations, April 2024.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback, June 2022.

XuanLong Nguyen, Martin J Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

OpenAI. Gpt-4 technical report. `https://openai.com/research/gpt-4`, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, December 2022.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal Language Modeling. In *The Twelfth International Conference on Learning Representations*, October 2023.

Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. Conformal Nucleus Sampling. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 27–34, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.3.

Jan Heinrich Reimer, Sebastian Schmidt, Maik Fröbe, Lukas Gienapp, Harrisen Scells, Benno Stein, Matthias Hagen, and Martin Potthast. The Archive Query Log: Mining Millions of Search Result Pages of Hundreds of Search Engines from 25 Years of Web Archives. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 2848–2860, New York, NY, USA, July 2023. Association for Computing Machinery. ISBN 978-1-4503-9408-6. doi: 10.1145/3539618.3591890.

Glenn Shafer and Vladimir Vovk. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. ISSN 1533-7928.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the Bregman divergence: A unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, October 2012. ISSN 1572-9052. doi: 10.1007/s10463-011-0343-8.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal Prediction Under Covariate Shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023.

Dennis Ulmer, Chrysoula Zerva, and Andre Martins. Non-Exchangeable Conformal Language Generation with Nearest Neighbors. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1909–1929, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

HuaJun Xi, Kangdao Liu, Hao Zeng, Wenguang Sun, and Hongxin Wei. Exploring the Noise Robustness of Online Conformal Prediction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025.

Yu-Jie Zhang, Zhen-Yu Zhang, Peng Zhao, and Masashi Sugiyama. Adapting to Continuous Covariate Shift via Online Density Ratio Estimation. *Advances in Neural Information Processing Systems*, 36:29074–29113, December 2023.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations*, October 2023.

A The Use of LLMs

In this work, LLMs were solely used for text polishing to enhance the clarity and readability of the manuscript. LLMs played no role in generating research ideas, problem formulations, proofs, theorems, algorithms, experiments, results, figures, or evaluations. All content produced or refined by LLMs was meticulously reviewed and validated by the authors to ensure its accuracy and consistency with the intended meaning. The technical contributions and intellectual work presented in this study are entirely the authors' own.

B Related Work

Hallucination and Its Mitigation Hallucination in LLMs represents a critical research challenge, driving extensive efforts to improve their factuality. Based on the stage of application, these efforts can generally be divided into two categories: inference-level and training-level methods. At the inference stage, a common approach is retrieval-augmented generation (RAG), which grounds model responses in external knowledge sources (Lewis et al., 2020; Fan et al., 2024; Jiang et al., 2023). Training-level strategies include reinforcement learning from human feedback (RLHF) (Bai et al., 2022; Ouyang et al., 2022), supervised fine-tuning with factual supervision (Lin et al., 2022; Nakano et al., 2022), and factuality-oriented decoding techniques (Chuang et al., 2023; Lee et al., 2022).

In addition to these heuristic approaches, recent research has introduced methods with statistical guarantees. For instance, Mohri & Hashimoto (2024) applied conformal prediction to LLMs to provide marginal guarantees on factuality, while Cherian et al. (2024) extended this framework to group-wise guarantees using conditional conformal prediction (Gibbs et al., 2025). However, these methods are built on the assumption of exchangeability between calibration and test data—an assumption that is often violated in real-world covariate shifts. This highlights the need for more robust methods that are better suited to practical scenarios.

Conformal Prediction Conformal prediction (CP) provides a formal framework for constructing prediction sets with guaranteed coverage (Shafer & Vovk, 2008; Vovk et al., 2005). Recently, there has been growing interest in applying CP to calibrate LLM outputs and improve their reliability (Campos et al., 2024). At the response level, Kumar et al. (2023) and Quach et al. (2023) leverage CP to identify low-confidence outputs, enhancing the reliability of model predictions. At the token level, Ulmer et al. (2024) and Ravfogel et al. (2023) employ CP-guided decoding to improve text quality. While these methods have demonstrated empirical success in improving reliability, they often fail to produce a single response with guaranteed factuality—an outcome that is typically more practical and desirable than generating a prediction set containing potential valid outputs.

Another important research direction involves extending CP to online settings, which better reflect real-world sequential applications. Gibbs & Candes (2021) introduced adaptive conformal inference (ACI) to maintain coverage under distribution shifts. Building on this, Gibbs & Candès (2024) proposed adaptive step-size tuning to improve ACI's robustness. More recently, Areces et al. (2025) and Bhatnagar et al. (2023) developed advanced online learning algorithms that guarantee coverage at a finer granularity, rather than averaging coverage over the entire time horizon. However, most online CP methods rely on the assumption of immediate access to ground-truth labels for test data following predictions—an assumption that is not feasible in the context of hallucination mitigation for LLMs, where feedback on output correctness is typically unavailable. Consequently, existing online CP frameworks are unsuitable for this problem, highlighting the need for new approaches designed to address these constraints.

C Omitted Algorithm Details

In this section, we outline the approach to minimizing dynamic regret, as defined in Equation 14, using the online ensemble framework proposed by Zhang et al. (2023). At a high level, the framework maintains a pool of experts, where each expert models a density ratio estimator over its designated lifetime. At each time step, the predictions from all active experts are aggregated to form a global model, which provides the final density ratio estimation.

Algorithm 1 CoFact's Online DRE Framework, adapted from Zhang et al. (2023)

```
Require: Calibration data $D_0 = \{Z_i\}_{i=1}^n$, number of time steps $T$
 1: Initialize the set of lifetime length list $\mathcal{C} = [1, 2, 4, ...\lceil \log_2 T \rceil]$
 2: Initialize the active set of experts $\mathcal{A}$ with $|\mathcal{C}|$ initialized experts
 3: for $t = 1, \ldots, T$ do
 4:     for $L \in \mathcal{C}$ do
 5:         if $t \equiv 0 \mod L$ then
 6:             Reinitialize the expert (its $\theta$, $\varepsilon$ and $v$) corresponding to the lifetime length $L$, i.e., $\mathcal{A}[\log_2 L]$
 7:     for $\mathcal{E}_i \in \mathcal{A}$ do
 8:         Update $p_{t,i}$ using $\varepsilon_{t-1,i}$ and $v_{t-1,i}$
 9:     Aggregate the global model $\hat{\theta}_t$
10:     for $\mathcal{E}_i \in \mathcal{A}$ do
11:         Update the parameters of $\mathcal{E}_i$, i.e., $\hat{\theta}_{t+1,i}$
12:         Update the potential $v_{t,i}$ and step size $\varepsilon_{t,i}$
```

The overall algorithm is detailed in Algorithm 1, which consists of three main steps: active-set update (lines 3–6), model aggregation (lines 7–9), and expert update (lines 10–12). Below, we provide an overview of each step.

Active-set update. The algorithm maintains a set of experts, each assigned a lifetime length chosen geometrically as $2^0, 2^1, 2^2, \ldots$. At each time step t, the algorithm checks if any expert's lifetime has expired. If so, the expired expert is re-initialized with updated parameters, including the model parameter $\hat{\theta}_{t,i}$, the potential $v_{t,i}$, and the step size $\epsilon_{t,i}$. Specifically, the initialization procedure is as follows:

- **Model parameter initialization.** The model parameter $\hat{\theta}_{t,i}$ is initialized to the current global model: $\hat{\theta}_{t,i} = \hat{\theta}_t$.
- **Potential initialization.** The potential $v_{t,i}$ is initialized as $v_{t,i} = 1/T$.
- **Step size initialization.** The step size $\epsilon_{t,i}$ is initialized as $\epsilon_{t,i} = \min\{1/2, \sqrt{\ln T}\}$.

Model aggregation. After updating the active experts, the algorithm aggregates their predictions to form a global model. This global model is then used to make predictions for the current test sample Z_{n+t}. For each expert, a "potential" $v_{t,i}$ is maintained to reflect its historical performance, while a step size $\epsilon_{t,i}$ controls the update of this potential. The weights and the global model are computed as follows:

$$p_{t,i} = \frac{\epsilon_{t-1,i} v_{t-1,i}}{\sum_{i \in \mathcal{A}} \epsilon_{t-1,i} v_{t-1,i}}, \quad \text{and} \quad \hat{\theta}_t = \sum_{i \in \mathcal{A}} p_{t,i} \hat{\theta}_{t,i}.$$

Expert update. Once the global model $\hat{\theta}_t$ is obtained, each active expert $\mathcal{E}_i$ is updated using the newly arrived test sample Z_{n+t}. The expert update consists of two components: model parameter update and updates to the potential and step size.

1. **Model parameter update.** The model parameter $\hat{\theta}_{t,i}$ is updated using the online Newton step (ONS) method (Hazan et al., 2007; Agarwal et al., 2006), which incorporates second-order information to achieve efficient and adaptive updates. The update rule is given by:

$$\hat{\theta}_{t+1,i} = \Pi_{\Theta}^{A_{t,i}} \left[\hat{\theta}_{t,i} - \gamma A_{t,i}^{-1} \nabla \hat{L}_t(\hat{\theta}_{t,i}) \right],$$

where $A_{t,i}$ is the accumulated second-order matrix defined as:

$$A_{t,i} = \lambda I + \sum_{\tau = s_i}^{t} \nabla \hat{L}_\tau(\hat{\theta}_{\tau,i}) \nabla \hat{L}_\tau(\hat{\theta}_{\tau,i})^\top,$$

and s_i denotes the last initialization time step of expert $\mathcal{E}_i$. The term $\Pi_{\Theta}^{A_{t,i}}$ represents the projection of the updated parameter onto the feasible set Θ, with the projection performed under the norm induced by the matrix $A_{t,i}$. This ensures that the updated parameter remains within the allowable parameter space.

2. **Potential and step size update.** The potential $v_{t,i}$ and step size $\epsilon_{t,i}$ are updated to reflect the expert's performance. First, we define the term $m_{t,i}$, which captures the performance gap between the expert $\hat{\theta}_{t,i}$ and the global model $\hat{\theta}_t$ over the linearized loss:

$$m_{t,i} = \frac{\langle \nabla \hat{L}_t(\hat{\theta}_t), \hat{\theta}_t - \hat{\theta}_{t,i} \rangle}{SR}.$$

Using $m_{t,i}$, the updates are performed as follows:

- **Potential update.** The potential $v_{t,i}$ is updated as:

$$v_{t,i} = v_{t-1,i} \cdot (1 + \epsilon_{t-1,i} m_{t,i})^{\frac{\epsilon_{t,i}}{\epsilon_{t-1,i}}}.$$

- **Step size update.** The step size $\epsilon_{t,i}$ is updated as:

$$\epsilon_{t,i} = \min\left\{\frac{1}{2}, \sqrt{\frac{\ln T}{1 + \sum_{\tau=s_i}^{t} m_{\tau,i}^2}}\right\}.$$

D Theorem and Proof

D.1 Useful Lemma

Lemma 1 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in [l_{\text{low}}, l_{\text{low}} + L]$ almost surely. Define the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[\bar{X}_n]$. Then for any $\delta > 0$, with probability at least $1 - \delta$,*

$$|\bar{X}_n - \mu| \le L \cdot \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{21}$$

Lemma 2 (Azuma–Hoeffding Inequality)**.** *Let $\{M_t\}_{t=1}^n$ be a martingale difference sequence with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^n$, i.e.,*

$$\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] = 0, \quad \forall\, t = 1, \ldots, n.$$

Suppose the differences are bounded almost surely by

$$|M_t| \le c, \quad \forall\, t = 1, \ldots, n.$$

Define the partial average $B_n = \frac{1}{n}\sum_{t=1}^n M_t$. Then for any $\delta > 0$ with probability at least $1 - \delta$, we have

$$|B_n| \le c\sqrt{\frac{2}{n}\log(2/\delta)}. \tag{22}$$

Lemma 3 (Theorem 1 in Zhang et al. (2023))**.** *Suppose Assumptions 2 and 3 hold. Denote $[z]_+ = \max\{z, 0\}$. Let d be the dimension of the parameter space Θ. Then, for any density ratio estimator $\hat{r}_t(Z) = h(Z;\theta) \in \mathcal{H}_r$, the empirical estimation error is bounded by*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{x\sim D_0(x)}\left[|r_t^*(x) - \hat{r}_t(x)|\right] \le \sqrt{\frac{4}{\mu T}\left[\sum_{t=1}^{T}\tilde{L}_t^{\psi}(\hat{r}_t) - \sum_{t=1}^{T}\tilde{L}_t^{\psi}(r_t^*)\right]_+} + \mathcal{O}\left(\frac{\sqrt{d\log(T/\delta)}}{\mu\sqrt{n}}\right), \tag{23}$$

provided that $h(Z,\theta)$ is bounded for any $Z \in \mathcal{Z}$ and $\theta \in \Theta$ and Lipschitz continuous.

Lemma 4 (Theorem 2 in Zhang et al. (2023))**.** *Suppose Assumptions 2 and 3 hold. Then, with probability at least $1-\delta$, the dynamic regret of the density ratio estimator sequence $\{\hat{r}_t\}_{t=1}^T$ learned from Algorithm 1 is bounded by*

$$\sum_{t=1}^{T}\tilde{L}_t^{\psi}(\hat{r}_t) - \sum_{t=1}^{T}\tilde{L}_t^{\psi}(r_t^*) \le \tilde{\mathcal{O}}\left(\max\left\{T^{\frac{1}{3}}V_T^{\frac{2}{3}}, 1\right\} + \frac{T}{n}\right), \tag{24}$$

when the parameters are set as $\gamma = 3(1+\beta)$ and $\lambda = 1$. In the above, $V_T = \sum_{t=2}^{T}\|\mathcal{D}_t(\mathbf{x}) - \mathcal{D}_{t-1}(\mathbf{x})\|_1$ measures the variation of input densities. $\beta = \exp(SR)$ represents the maximum value of the estimated density ratio $\hat{r}_t$.

Corollary 2. *Suppose Assumption 2 and 3 hold. Then, with probability at least $1-\delta$, the dynamic regret of the density ratio estimator $\hat{r}_t(\mathbf{x}) = \exp(-\phi(\mathbf{z})^\top \hat{\theta}_t)$ is bounded by*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathbf{x}\sim D_0}\left[|r_t^*(\mathbf{x}) - \hat{r}_t(\mathbf{x})|\right] \le \tilde{O}\left(n^{-\frac{1}{2}} + \max\left\{T^{-\frac{1}{3}}V_T^{\frac{1}{3}}, T^{-\frac{1}{2}}\right\}\right). \tag{25}$$

when the parameters are set as $\gamma = 3(1+\beta)$ and $\lambda = 1$. In the above, $V_T = \sum_{t=2}^{T} \|\mathcal{D}_t(\mathbf{x}) - \mathcal{D}_{t-1}(\mathbf{x})\|_1$ measures the variation of input densities.

D.2 The proof of Theorem 1

For clarity, let $\mathcal{E}_t$ denote the event that there exists any hallucinated claim in the prediction set obtained by true density ratio using at time step t, i.e.,

$$\mathcal{E}_t = \Big\{\exists\, C_{n+t,i} \in F(\mathbf{C}_{n+t}) \text{ such that } W_{n+t,i} = 0\Big\}. \tag{26}$$

Similarly, we denote the event that there exists any hallucinated claim in the prediction set obtained by estimated density ratio at time step t as $\hat{\mathcal{E}}_t$, i.e.,

$$\hat{\mathcal{E}}_t = \Big\{\exists\, C_{n+t,i} \in \hat{F}(\mathbf{C}_{n+t}) \text{ such that } W_{n+t,i} = 0\Big\}. \tag{27}$$

As a result, we have $\mathrm{err}_t = \mathbb{1}[\mathcal{E}_t]$ and $\widehat{\mathrm{err}}_t = \mathbb{1}[\hat{\mathcal{E}}_t]$.

Lemma 5. *The empirical error rate $\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathrm{err}}_t$ concentrates around its expectation $\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}[\hat{\mathcal{E}}_t]$. Specifically, for any $\delta > 0$, with probability at least $1-\delta$,*

$$\left|\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathrm{err}}_t - \frac{1}{T}\sum_{t=1}^{T}\mathbb{P}[\hat{\mathcal{E}}_t]\right| \le \sqrt{\tfrac{2}{T}\log(2/\delta)}. \tag{28}$$

Proof. Define the centered random variables

$$Y_t = \mathbb{1}[\hat{\mathcal{E}}_t] - \mathbb{P}[\hat{\mathcal{E}}_t], \quad t = 1, \dots, T.$$

Clearly, $|Y_t| \le 1$. Let $\mathcal{F}_{t-1} = \sigma(D_0, D_1, \dots, D_{t-1})$ be the filtration generated by the calibration set and past test samples. We first verify that $\{Y_t\}_{t=1}^{T}$ forms a martingale difference sequence with respect to $\{\mathcal{F}_t\}_{t=0}^{T}$.

Since $(\mathbf{C}_{n+t}, \mathbf{W}_{n+t})$ is independent of $\mathcal{F}_{t-1}$ and $\hat{F}$ is $\mathcal{F}_{t-1}$-measurable, it follows that

$$\mathbb{E}[\mathbb{1}[\hat{\mathcal{E}}_t] \mid \mathcal{F}_{t-1}] = \mathbb{P}(\hat{\mathcal{E}}_t \mid \mathcal{F}_{t-1}) = \mathbb{P}(\hat{\mathcal{E}}_t|\hat{F}).$$

Consequently,

$$\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[\mathbb{1}[\hat{\mathcal{E}}_t] - \mathbb{P}[\hat{\mathcal{E}}_t] \mid \mathcal{F}_{t-1}] = \mathbb{P}(\hat{\mathcal{E}}_t|\hat{F}) - \mathbb{P}(\hat{\mathcal{E}}_t|\hat{F}) = 0.$$

Thus, $\{Y_t\}$ is a martingale difference sequence bounded in $[-1, 1]$.

By applying Azuma–Hoeffding's inequality (Lemma 2), we proof the lemma. □

Lemma 6. *Suppose Assumption 1 holds. Let $\hat{r}_t$ and r_t^* be the estimated and true density ratios at time step t, respectively, and let Z_{n+t} be a test sample drawn independently at time step t. Then, the following inequality holds with probability at least $1-\delta$:*

$$\frac{1}{T}\sum_{t=1}^{T}|\hat{r}_t(Z_{n+t}) - r_t^*(Z_{n+t})| \le \frac{B}{T}\sum_{t=1}^{T}\mathbb{E}_{Z\sim\mathcal{D}_0}\left[|r_t^*(Z) - \hat{r}_t(Z)|\right] + \sqrt{\frac{2}{T}\log\left(\frac{2}{\delta}\right)}\cdot\beta', \tag{29}$$

where β' is a bound on the differences $|\hat{r}_t(z) - r_t^*(z)|$ for all z.

Proof. We first define the centered variable for each time step t:

$$U_t = |\hat{r}_t(Z_{n+t}) - r_t^*(Z_{n+t})| - \mathbb{E}_{Z\sim\mathcal{D}_t}\left[|\hat{r}_t(Z) - r_t^*(Z)|\right]. \tag{30}$$

Given the independence between $\hat{r}_t$ and the test sample Z_{n+t}, and the fact that $\hat{r}_t$ is measurable with respect to $\mathcal{F}_{t-1} = \sigma(D_0, D_1, \ldots, D_{t-1})$, we can show that $\{U_t\}_{t=1}^T$ forms a martingale difference sequence with respect to the filtration $\{\mathcal{F}_t\}_{t=1}^T$ similar to the proof of Lemma 5. Applying the Azuma-Hoeffding inequality, we obtain that with probability at least $1-\delta$,

$$\begin{aligned}&\left|\frac{1}{T}\sum_{t=1}^{T}|\hat{r}_t(Z_{n+t}) - r_t^*(Z_{n+t})| - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{Z\sim\mathcal{D}_t}\left[|\hat{r}_t(Z) - r_t^*(Z)|\right]\right|\\ &\leq \sqrt{\frac{2}{T}\log\left(\frac{2}{\delta}\right)}\cdot\beta',\end{aligned} \tag{31}$$

Next, we bound the expected difference under the true data distribution $\mathcal{D}_t$, which can be related to the initial distribution $\mathcal{D}_0$:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{Z\sim\mathcal{D}_t}\left[|\hat{r}_t(Z) - r_t^*(Z)|\right] &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{Z\sim\mathcal{D}_0}\left[|r_t^*(Z)(\hat{r}_t(Z) - r_t^*(Z))|\right]\\ &\leq \frac{B}{T}\sum_{t=1}^{T}\mathbb{E}_{Z\sim\mathcal{D}_0}\left[|r_t^*(Z) - \hat{r}_t(Z)|\right],\end{aligned} \tag{32}$$

where B is a bound on $r_t^*(Z)$.

Combining Equation 31 and Equation 32, we prove the lemma. □

Lemma 7. *Suppose Assumption 1 holds. Given the hypothesis space $\mathcal{H}_r$ satisfying Assumption 2, let $\beta' = \max_{r\in\mathcal{H}_r, z\in\mathcal{Z}}|r(z) - r_t^*(z)|$ and $G_h = \max_{Z\in\mathcal{Z},\theta\in\Theta}\|\nabla h(Z,\theta)\|_2$. For any sequence of density ratio estimators $\{\hat{r}_t\}_{t=1}^T$ and corresponding true density ratio $\{r_t^*\}_{t=1}^T$ under distribution $\mathcal{D}_0$, the following bound holds with probability at least $1-\delta$:*

$$\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{Z\sim\mathcal{D}_0}[|\hat{r}_t(Z) - r_t^*(Z)|] - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{Z\sim D_0}[|\hat{r}_t(Z) - r_t^*(Z)|]\right| \leq 2\beta'\sqrt{\frac{d\log\left(\frac{6SG_hT}{\delta}\right)}{2n}} + \frac{2}{T} \tag{33}$$

Proof. To establish the bound, we consider the discrepancy between the empirical and expected absolute errors across a sequence of estimators. The primary challenge arises from the dependence between the estimators $\hat{r}_t$ and their parameters $\hat{\theta}_t$, which precludes the direct application of Hoeffding's inequality. To navigate this, we employ a two-step approach involving Hoeffding's inequality and the covering number theory.

Firstly, for a fixed model $r\in\mathcal{H}_r$ and a specific time $t\in[T]$, define:

$$g_r(Z) = |r(Z) - r_t^*(Z)| \tag{34}$$

and let $U_i = \mathbb{E}_{Z\sim\mathcal{D}_0}[g_r(Z)] - g_r(Z_i)$, where Z_i are i.i.d. samples from $\mathcal{D}_0$. Since g_r is independent of D_0, the variables U_i are independent and bounded by $[-\beta', \beta']$. By Hoeffding's inequality, we have:

$$\left|\mathbb{E}_{Z\sim\mathcal{D}_0}[g_r(Z)] - \frac{1}{n}\sum_{i=1}^{n}g_r(Z_i)\right| \leq \beta'\sqrt{\frac{\log(2/\delta)}{2n}} \tag{35}$$

with probability at least $1-\delta$.

Secondly, we extend the analysis to the entire hypothesis space $\mathcal{H}_r$ using the concept of covering numbers. Define $\mathcal{N}(\mathcal{H}_r, \epsilon, \|\cdot\|_\infty)$ as the ϵ-covering number of $\mathcal{H}_r$. By applying a union bound over all models in an ϵ-net of $\mathcal{H}_r$, we obtain:

$$\left|\mathbb{E}_{Z\sim\mathcal{D}_0}[g_{r'}(Z)] - \frac{1}{n}\sum_{i=1}^{n}g_{r'}(Z_i)\right| \leq \beta'\sqrt{\frac{\log\left(\frac{2\mathcal{N}(\mathcal{H}_r,\epsilon,\|\cdot\|_\infty)}{\delta}\right)}{2n}} \tag{36}$$

for all $r' \in \mathcal{N}(\mathcal{H}_r, \epsilon, \|\cdot\|_\infty)$, with probability at least $1-\delta$.

To establish a comprehensive bound on the difference between the expected and empirical measures of risk over the hypothesis space $\mathcal{H}_r$, we decompose the discrepancy for any model $r \in \mathcal{H}_r$ into three terms related to approximation, estimation, and covering errors. This decomposition allows us to systematically address each source of error and apply probabilistic bounds accordingly.

First, we define the discrepancy for a fixed model r and its approximation r' from the ϵ-net of $\mathcal{H}_r$:

$$\begin{aligned}
&\left|\mathbb{E}_{Z\sim D_0}[g_r(Z)] - \frac{1}{n}\sum_{i=1}^n g_r(Z_i)\right| \\
&\le \underbrace{|\mathbb{E}_{Z\sim D_0}[g_r(Z)] - \mathbb{E}_{Z\sim D_0}[g_{r'}(Z)]|}_{\text{Term}(a)} + \underbrace{\left|\mathbb{E}_{Z\sim D_0}[g_{r'}(Z)] - \frac{1}{n}\sum_{i=1}^n g_{r'}(Z_i)\right|}_{\text{Term}(b)} \qquad (37) \\
&\quad + \underbrace{\left|\frac{1}{n}\sum_{i=1}^n g_{r'}(Z_i) - \frac{1}{n}\sum_{i=1}^n g_r(Z_i)\right|}_{\text{Term}(c)}.
\end{aligned}$$

We analyze each term separately:

- **Term (a)** $|\mathbb{E}_{Z\sim D_0}[g_r(Z)] - \mathbb{E}_{Z\sim D_0}[g_{r'}(Z)]|$ captures the approximation error due to using r' instead of r. By the properties of the ϵ-net, this term is bounded by ϵ.
- **Term (b)** $\left|\mathbb{E}_{Z\sim D_0}[g_{r'}(Z)] - \frac{1}{n}\sum_{i=1}^n g_{r'}(Z_i)\right|$ represents the estimation error for the approximating model r'. Using Hoeffding's inequality and considering the covering number of $\mathcal{H}_r$, this term is bounded by :

$$\beta'\sqrt{\frac{\log\left(\frac{2\mathcal{N}(\mathcal{H}_r,\epsilon,\|\cdot\|_\infty)}{\delta}\right)}{2n}}. \qquad (38)$$

  with probability at least $1-\delta$.
- **Term (c)** $\left|\frac{1}{n}\sum_{i=1}^n g_{r'}(Z_i) - \frac{1}{n}\sum_{i=1}^n g_r(Z_i)\right|$ is also bounded by ϵ, similar to term (a), due to the ϵ-closeness of r and r'.

Adding these terms, the total bound for any model $r \in \mathcal{H}_r$ is:

$$\left|\mathbb{E}_{Z\sim D_0}[g_r(Z)] - \frac{1}{n}\sum_{i=1}^n g_r(Z_i)\right| \le 2\epsilon + \beta'\sqrt{\frac{\log\left(\frac{2\mathcal{N}(\mathcal{H}_r,\epsilon,\|\cdot\|_\infty)}{\delta}\right)}{2n}}. \qquad (39)$$

By setting $\epsilon = \frac{1}{T}$ and summing over all $t \in [T]$, the final aggregated bound for the entire sequence of estimators under consideration becomes:

$$\left|\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{Z\sim D_0}[g_{\hat{r}_t}(Z)] - \frac{1}{T}\sum_{t=1}^T \frac{1}{n}\sum_{i=1}^n g_{\hat{r}_t}(Z_i)\right| \le \frac{2}{T} + \beta'\sqrt{\frac{\log\left(\frac{2\mathcal{N}(\mathcal{H}_r,1/T,\|\cdot\|_\infty)}{\delta}\right)}{2n}}. \qquad (40)$$

Next, we focus on bounding the covering number of the hypothesis space $\mathcal{H}_r$. Since we parameterize the density ratio functions in $\mathcal{H}_r$ using a parametric model $h(\mathbf{x},\theta)$ with parameters θ in a bounded set Θ, we can relate the covering number of $\mathcal{H}_r$ to that of Θ.

Let $\theta, \theta' \in \Theta$ be the parameters corresponding to the two density ratio functions $r, r' \in \mathcal{H}_\theta$. We can show that for any $\|\theta-\theta'\|_2 \le \epsilon$, the following inequality holds:

$$\|r - r'\|_\infty = \max_{Z\in\mathcal{Z}} |r(Z,\theta) - r(Z,\theta')| \le G_h\|\theta-\theta'\|_2,$$

where $G_h = \max_{Z\in\mathcal{Z},\theta\in\Theta} \|\nabla h(Z,\theta)\|_2$ is the Lipschitz continuity constant of h.

As a result, we can bound the covering number of $\mathcal{H}_r$ in terms of $\|\cdot\|_\infty$ by the covering number of Θ in terms of $\|\cdot\|_2$. Specifically, we have:

$$\mathcal{N}(\mathcal{H}_r, 1/T, \|\cdot\|_\infty) \leq \mathcal{N}(\Theta, 1/(G_h T), \|\cdot\|_2).$$

Given that the parameter space Θ is essentially a L_2-ball with radius S, its covering number is bounded by $(3S/\epsilon)^d$. Therefore, choosing $\epsilon = 1/(G_h T)$, we obtain:

$$\mathcal{N}(\Theta, 1/(G_h T), \|\cdot\|_2) \leq (3SG_h T)^d.$$

Combining these results, we conclude:

$$\mathcal{N}(\mathcal{H}_\theta, 1/T, \|\cdot\|_\infty) \leq (3SG_h T)^d.$$

Substituting this bound into our earlier expression, we complete the proof.

□

Lemma 8. *Suppose Assumption 1 holds. Given the hypothesis space $\mathcal{H}_r$ satisfying Assumption 2 and divergence function ψ satisfying Assumption 3. For any sequence of density ratio estimators $\{\hat{r}_t\}_{t=1}^T$ and corresponding true density ratios $\{r_t^*\}_{t=1}^T$ under distribution $\mathcal{D}_0$, the following bound holds with probability at least $1-\delta$:*

$$\frac{1}{T}\sum_{t=1}^{T} |\hat{r}_t(Z_{n+t}) - r_t^*(Z_{n+t})| \leq \tilde{\mathcal{O}}\left(n^{-\frac{1}{2}} + \max\{T^{-\frac{1}{3}}V_T^{\frac{1}{3}}, T^{-\frac{1}{2}}\}\right) \tag{41}$$

Proof. The proof is straightforward by combining the results from Lemma 6, Lemma 7, and Corollary 2.

□

Lemma 9. *Suppose assumptions 1, 2 and 3 hold. Let $\{\mathcal{E}_t\}_{t=1}^T$ and $\{\hat{\mathcal{E}}_t\}_{t=1}^T$ represent two sequences of events defined in Equation 26 and 27. The difference in their average probabilities satisfies:*

$$\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}[\mathcal{E}_t] - \frac{1}{T}\sum_{t=1}^{T}\mathbb{P}[\hat{\mathcal{E}}_t]\right| \leq \tilde{O}\left(n^{-\frac{1}{2}} + \max\left\{T^{-\frac{1}{3}}V_T^{\frac{1}{3}}, T^{-\frac{1}{2}}\right\}\right). \tag{42}$$

Proof. Define the true weighted cumulative distribution function (CDF) at time step t as:

$$\Psi_t(v) = \sum_{i\in[n]\cup\{n+t\}} w_t^*(Z_i)\mathbb{1}[v_i \leq v],$$

where $w_t^*(Z_i)$ are the true weights derived from the true density ratio r_t^*, and v_i denotes the value of $V(\mathbf{C}_i, \mathbf{W}_i)$. For simplicity, v_{n+t} is set to ∞.

Similarly, the estimated weighted CDF at time step t is:

$$\hat{\Psi}_t(v) = \sum_{i\in[n]\cup\{n+t\}} \hat{w}_t(Z_i)\mathbb{1}[v_i \leq v],$$

where $\hat{w}_t(Z_i) = \frac{\hat{r}_t(Z_i)}{\sum_{j\in[n]\cup\{n+t\}}\hat{r}_t(Z_j)}$ are the estimated weights based on the estimated density ratio $\hat{r}_t$.

Due to the nested property, i.e., the size of $\hat{F}_t$ non-decreasing w.r.t $\hat{\tau}_t$, the following equation holds:

$$\mathcal{E}_t = \{V(\mathbf{C}_{n+t}, \mathbf{W}_{n+t}) > \tau_t\}, \quad \hat{\mathcal{E}}_t = \{V(\mathbf{C}_{n+t}, \mathbf{W}_{n+t}) > \hat{\tau}_t\},$$

where τ_t and $\hat{\tau}_t$ are thresholds derived from Ψ_t and $\hat{\Psi}_t$, respectively. Using these definitions, we can express the difference in average probabilities as:

$$\begin{aligned}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}[\mathcal{E}_t]-\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}[\hat{\mathcal{E}}_t]\right| &= \left|\frac{1}{T}\sum_{t=1}^{T}\Psi_t(\hat{\tau}_t)-\Psi_t(\tau_t)\right| \\ &= \left|\frac{1}{T}\sum_{t=1}^{T}\Psi_t(\hat{\tau}_t)-\Psi_t(\tau_t)+\Psi_t(\tau_t)-\hat{\Psi}_t(\hat{\tau}_t)\right| \\ &= \left|\frac{1}{T}\sum_{t=1}^{T}\Psi_t(\hat{\tau}_t)-\hat{\Psi}_t(\hat{\tau}_t)\right|.\end{aligned}$$

The last equality holds because $\Psi_t(\tau_t)=\hat{\Psi}_t(\hat{\tau}_t)=1-\alpha$, by definition of the thresholds τ_t and $\hat{\tau}_t$. Expanding $\Psi_t(\hat{\tau}_t)$ and $\hat{\Psi}_t(\hat{\tau}_t)$, we have:

$$\left|\frac{1}{T}\sum_{t=1}^{T}\Psi_t(\hat{\tau}_t)-\hat{\Psi}_t(\hat{\tau}_t)\right| = \left|\frac{1}{T}\sum_{t=1}^{T}\sum_{i\in[n]\cup\{n+t\}}\mathbb{1}[v_i\le\hat{\tau}_t]\left(w_t^*(Z_i)-\hat{w}_t(Z_i)\right)\right|.$$

Substituting $w_t^*(Z_i)$ and $\hat{w}_t(Z_i)$, we obtain:

$$\begin{aligned}&\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{i\in[n]\cup\{n+t\}}\mathbb{1}[v_i\le\hat{\tau}_t]\left(\frac{r_t^*(Z_i)}{\sum_{j\in[n]\cup\{n+t\}}r_t^*(Z_j)}-\frac{\hat{r}_t(Z_i)}{\sum_{j\in[n]\cup\{n+t\}}\hat{r}_t(Z_j)}\right)\right| \\ \le &\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{i\in[n]\cup\{n+t\}}\mathbb{1}[v_i\le\hat{\tau}_t]\left(\frac{r_t^*(Z_i)\cdot\sum_{j\in[n]\cup\{n+t\}}\hat{r}_t(Z_j)-\hat{r}_t(Z_i)\sum_{j\in[n]\cup\{n+t\}}r_t^*(Z_j)}{\sum_{j\in[n]\cup\{n+t\}}r_t^*(Z_j)\cdot\sum_{j\in[n]\cup\{n+t\}}\hat{r}_t(Z_j)}\right)\right|.\end{aligned}$$

Simplifying the numerator of the fraction, let:

$$\text{Term}_i = r_t^*(Z_i)\cdot\sum_{j\in[n]\cup\{n+t\}}\hat{r}_t(Z_j)-\hat{r}_t(Z_i)\cdot\sum_{j\in[n]\cup\{n+t\}}r_t^*(Z_j).$$

Expanding Term_i, we have:

$$\text{Term}_i = r_t^*(Z_i)\cdot\sum_{j\in[n]\cup\{n+t\}}(\hat{r}_t(Z_j)-r_t^*(Z_j))+(r_t^*(Z_i)-\hat{r}_t(Z_i))\cdot\sum_{j\in[n]\cup\{n+t\}}r_t^*(Z_j).$$

Summing over $i\in[n]\cup\{n+t\}$, we bound $|\text{Term}_i|$ as:

$$\sum_{i\in[n]\cup\{n+t\}}|\text{Term}_i|\le\sum_{i\in[n]\cup\{n+t\}}r_t^*(Z_i)\cdot\sum_{j\in[n]\cup\{n+t\}}|\hat{r}_t(Z_j)-r_t^*(Z_j)|+\sum_{i\in[n]\cup\{n+t\}}|r_t^*(Z_i)-\hat{r}_t(Z_i)|\cdot\sum_{j\in[n]\cup\{n+t\}}r_t^*(Z_j).$$

Combining terms and simplifying, we find:

$$\begin{aligned}&\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}[\mathcal{E}_t]-\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}[\hat{\mathcal{E}}_t]\right| \\ =&\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{i\in[n]\cup\{n+t\}}\mathbb{1}[v_i\le\hat{\tau}_t]\left(\frac{\text{Term}_i}{\sum_{j\in[n]\cup\{n+t\}}r_t^*(Z_j)\cdot\sum_{j\in[n]\cup\{n+t\}}\hat{r}_t(Z_j)}\right)\right| \\ \le&\frac{1}{T}\sum_{t=1}^{T}\sum_{i\in[n]\cup\{n+t\}}\frac{|\text{Term}_i|}{\sum_{j\in[n]\cup\{n+t\}}r_t^*(Z_j)\cdot\sum_{j\in[n]\cup\{n+t\}}\hat{r}_t(Z_j)} \\ \le&\frac{1}{T}\sum_{t=1}^{T}\frac{2}{\sum_{j\in[n]\cup\{n+t\}}\hat{r}_t(Z_j)}\cdot\sum_{j\in[n]\cup\{n+t\}}|\hat{r}_t(Z_j)-r_t^*(Z_j)|\end{aligned}$$

Using the assumption that $\hat{r}_t(Z_j)$ and $r_t^*(Z_j)$ are bounded, we further simplify:

$$\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}[\mathcal{E}_t]-\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}[\hat{\mathcal{E}}_t]\right|\leq 2\beta\left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{j\in[n]}|\hat{r}_t(Z_j)-r_t^*(Z_j)|+\frac{1}{n}\cdot\frac{1}{T}\sum_{t=1}^{T}|\hat{r}_t(Z_{n+t})-r_t^*(Z_{n+t})|\right].$$

Finally, bounding the terms using Corollary 2 and Lemma 8, we conclude the proof. □

Theorem 1. *Under the assumptions 1, 2 and 3, with probability at least $1-\delta$, the gap between the averaged hallucination rate over T time steps and the target level α is bounded as*

$$\left|\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathrm{err}}_t-\alpha\right|\leq\tilde{\mathcal{O}}\left(\max\left\{T^{-\frac{2}{3}}V_T^{\frac{2}{3}},T^{-\frac{1}{2}}\right\}+1/n\right) \tag{19}$$

when the parameter of the online ensemble is properly set. Here, $V_T=\sum_{t=2}^{T}\|\mathcal{D}_t(\mathbf{z})-\mathcal{D}_{t-1}(\mathbf{z})\|_1$ measures the variation of input densities and the notation $\tilde{\mathcal{O}}$ hides logarithmic factors of T and $1/\delta$.

Proof. The error decomposition is:

$$\left|\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathrm{err}}_t-\alpha\right|\leq\underbrace{\left|\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathrm{err}}_t-\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}[\hat{\mathcal{E}}_t]\right|}_{\text{term (a)}}+\underbrace{\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}[\hat{\mathcal{E}}_t]-\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}[\mathcal{E}_t]\right|}_{\text{term (b)}}+\underbrace{\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}[\mathcal{E}_t]-\alpha\right|}_{\text{term (c)}}. \tag{43}$$

Using Lemma 5, term (a) is bounded as:

$$\text{term (a)}\leq\tilde{\mathcal{O}}(T^{-\frac{1}{2}}).$$

Using Lemma 9, term (b) is bounded as:

$$\text{term (b)}\leq\tilde{\mathcal{O}}\left(n^{-\frac{1}{2}}+\max\left\{T^{-\frac{1}{3}}V_T^{\frac{1}{3}},T^{-\frac{1}{2}}\right\}\right).$$

By definition, $\mathbb{P}[\mathcal{E}_t]=\alpha$, so term (c) is zero:

$$\text{term (c)}=0.$$

Combining these bounds yields:

$$\left|\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathrm{err}}_t-\alpha\right|\leq\tilde{\mathcal{O}}\left(\max\left\{T^{-\frac{2}{3}}V_T^{\frac{2}{3}},T^{-\frac{1}{2}}\right\}+n^{-\frac{1}{2}}\right).$$

This completes the proof. □

E Omitted Experimental Details

E.1 Omitted Dataset Details

MedLFQA (Jeong et al. (2024)) MedLFQA is a long-form medical question-answering dataset that integrates several previously established benchmarks. Each prompt in the dataset is paired with responses generated by either an LLM or a human. To obtain sub-claims annotated with factuality labels, Cherian et al. (2024) first use GPT-3.5-Turbo to generate responses for the prompts and GPT-4o to parse these responses into self-contained sub-claims. Then, the factuality of each sub-claim is assessed by querying GPT-3.5-Turbo, which evaluates the claims based on the LLM or human-generated responses provided for the prompts.

WikiData (Cherian et al. (2024)) WikiData is constructed by first sampling names from Wikipedia and then querying GPT-3.5-Turbo with the prompt: *"Write me a short biography of [NAME]."*. After that, the generated biographies are parsed into self-contained sub-claims using GPT-4o. Factuality labels for these sub-claims are assigned using a variant of the FAcTScore procedure developed by Min et al. (2023). This process involves identifying relevant Wikipedia passages using the BM25 ranking function and incorporating them into the LLM prompt to determine whether the claims are supported.

WildChat+ To evaluate our method in a real-world dynamic setting, we construct a new dataset from WildChat Zhao et al. (2023), which features user-generated prompts in natural, uncontrolled environments. Since not all responses elicited by prompts in the dataset are suitable for the hallucination mitigation task, we first filter them by using GPT-4o-mini to identify whether the prompts can be answered using knowledge available on Wikipedia. For the filtered prompts, GPT-4o assigns relevant Wikipedia titles, and we retrieve the corresponding passages using the Wikipedia API. Finally, we apply the FAcTScore procedure, following Cherian et al. (2024), to annotate the factuality labels of the claims in the responses. Due to the high cost of annotation, we randomly sample 3250 prompts from the filtered prompts for annotation.

E.2 Omitted Details of Shift Simulation

Here, we describe the procedure for sampling Z_{n+t} from D_{test} to simulate covariate shifts. Each prompt is associated with metadata, which is used to define a feature vector $\mathbf{x}$. Prompts Z_{n+t} are sampled with probabilities proportional to $w(\mathbf{x}) = \exp(\mathbf{x}^T((1-\xi_t)\nu' + \xi_t\nu''))$, where ν' and ν'' are predefined weight vectors, and $\xi_t \in [0, 1]$ is a time-varying factor. Since both D_0 and D_{test} originate from the same underlying distribution, this resampling strategy effectively simulates a shift between the initial distribution $\mathcal{D}_0$ and the time-dependent test distribution $\mathcal{D}_t$.

For the MedLFQA dataset, the feature vector $\mathbf{x}$ is defined using five attributes, as detailed in Cherian et al. (2024): the length of the prompt, the length of the response, the mean log-probability of the response given the prompt, the standard error of the log-probability of the response given the prompt, and the dataset from which the prompt originates. ν' and ν'' are configured such that ν'' assigns higher weights to prompts with longer responses so that the sampling favors prompts with longer responses over time.

For the WikiData dataset, the feature vector $\mathbf{x}$ is constructed using the first, second, and third powers of the number of views received by the Wikipedia pages corresponding to each prompt, following Cherian et al. (2024). In this case, ν' and ν'' are chosen such that ν'' assigns higher weights to prompts associated with less-viewed Wikipedia pages.

The time-varying factor ξ_t is introduced to model four types of covariate shifts: Linear Shift (**Lin**), Square Shift (**Squ**), Sine Shift (**Sin**), and Bernoulli Shift (**Ber**). Each type captures distinct real-world scenarios in which data distributions evolve over time:

- **Linear Shift (Lin)**: ξ_t is defined as $\xi_t = t/T$, representing a gradual and continuous change in the environment over T time steps.
- **Square Shift (Squ)**: ξ_t alternates between 1 and 0 every M steps, where $2M$ defines the period length. To model a rapidly changing environment with periodic behavior, we set $M = \Theta(\sqrt{T})$.
- **Sine Shift (Sin)**: ξ_t follows a sinusoidal pattern, defined as $\xi_t = \sin(\pi t/M)$, where M represents the period length. Similar to the Square Shift, we set $M = \Theta(\sqrt{T})$.
- **Bernoulli Shift (Ber)**: ξ_t retains the value of ξ_{t-1} with probability $p \in [0, 1]$, and switches to $1-\xi_{t-1}$ with probability $1-p$. To simulate a rapidly changing environment, we set $p = \Theta(1/\sqrt{T})$.

E.3 Omitted Implementation Details

We preprocess the MedLFQA and WikiData datasets in accordance with the methodology described by Cherian et al. (2024). For the baseline methods, SCP and CondConf, we utilize the original code provided by Cherian et al. (2024). To implement our approach, we adapt the code from Zhang et al. (2023) to perform the online density ratio estimation procedure, using the same hyperparameter

Table 4: Comparison with the method handling covariate shift on MedLFQA dataset.

	Lin		Squ		Sin		Ber		Fix	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
Total	0.849±0.025	0.856±0.015	0.849±0.021	0.861±0.012	0.848±0.022	0.858±0.013	0.855±0.015	0.853±0.011	0.859±0.011	0.862±0.012
Partial	0.819±0.023	0.900±0.010	0.825±0.026	0.898±0.012	0.822±0.025	0.897±0.012	0.836±0.016	0.895±0.008	0.842±0.012	0.883±0.011
CoFact	0.895±0.026	0.715±0.031	0.897±0.022	0.718±0.030	0.894±0.018	0.715±0.031	0.900±0.019	0.713±0.027	0.904±0.014	0.713±0.027

Table 5: Comparison with the method handling covariate shift on WikiData dataset.

	Lin		Squ		Sin		Ber		Fix	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
Total	0.895±0.010	0.768±0.004	0.895±0.009	0.768±0.005	0.895±0.008	0.768±0.004	0.887±0.012	0.773±0.004	0.895±0.009	0.771±0.004
Partial	0.884±0.013	0.780±0.007	0.885±0.012	0.780±0.008	0.884±0.011	0.780±0.007	0.875±0.014	0.783±0.007	0.886±0.012	0.783±0.007
CoFact	0.896±0.010	0.748±0.006	0.895±0.009	0.748±0.006	0.895±0.008	0.748±0.006	0.897±0.008	0.749±0.006	0.896±0.010	0.753±0.007

configuration as specified in their original implementation. All experiments are conducted on a machine equipped with a 16-core Ultra 9 285H CPU.

F Additional Experimental Results

In this section, we present additional experimental results including comparisons with several oracle baselines and ablation studies of various components of our method.

F.1 Comparison with Method in Conformal Prediction under Covariate Shift

To provide a more comprehensive evaluation of CoFact, we compared it against existing approaches for handling covariate shifts by implementing the method proposed by Tibshirani et al. (2019), which specifically addresses covariate shift in conformal prediction. This method requires access to the entire set of test samples to train a fixed density ratio estimator between the training and test distributions. To thoroughly assess its performance, we implemented two variants of this method: (1) **Total**: This variant uses all test samples to estimate the density ratio and filter potentially hallucinated sub-claims; (2) **Partial**: This variant uses only the test samples observed during the first 500 time steps to estimate the density ratio and filter potentially hallucinated sub-claims for all time steps. It is important to emphasize that both variants are impractical in our online setting, as they require access to either the entire set of test samples or a substantial portion of them in advance, which violates the constraints of online learning.

In addition to the four shifting scenarios introduced in the manuscript, we also evaluated these variants under an additional scenario, **Fixed Shift (Fix)**, where the test distribution remains constant but differs from the training distribution—a setting similar to that adopted by Tibshirani et al. (2019).

The results of these experiments on the MedLFQA and WikiData datasets are presented in Table 4 and Table 5, respectively. From these results, we observe the following: First, The **Total** variant performs better than the **Partial** variant, indicating that access to all test samples improves density ratio estimation and enables the construction of more accurate factuality control. Second, CoFact consistently outperforms both **Total** and **Partial** variants across nearly all shift types and datasets in terms of factuality control. Even under fixed shifts, CoFact achieves superior or comparable performance to the baselines, demonstrating its robustness in handling both stable and dynamic covariate shifts.

F.2 Comparison with Method in Online Conformal Prediction

We also compare our method with the approach proposed by Gibbs & Candes (2021), which is designed for online conformal prediction settings where the test distribution may shift over time. However, unlike our setting, this method requires access to true labels immediately after making predictions at each time step. While this assumption is not feasible in our scenario, we implemented the method as a baseline for comparison.

We evaluated two update strategies from Gibbs & Candes (2021): **ACI-Simple**, defined by Equation (2) in the original paper, and **ACI-Momentum**, defined by Equation (3). Additionally, we varied the step size parameter γ over $\{0.0005, 0.001, 0.005, 0.01\}$ to analyze its impact on performance. The results on the MedQA and Wiki datasets are presented in Table 6 and Table 7, respectively.

Table 6: Comparison with the method handling online conformal prediction on MedLFQA dataset.

	Lin		Squ		Sin		Ber	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
ACI-Simple-0.0005	0.882±0.006	0.747±0.035	0.884±0.003	0.723±0.055	0.885±0.002	0.736±0.056	0.884±0.001	0.751±0.038
ACI-Momentum-0.0005	0.882±0.007	0.747±0.035	0.885±0.003	0.723±0.055	0.885±0.003	0.733±0.056	0.885±0.001	0.747±0.037
ACI-Simple-0.001	0.909±0.003	0.635±0.051	0.912±0.003	0.617±0.073	0.912±0.002	0.608±0.077	0.911±0.003	0.653±0.070
ACI-Momentum-0.001	0.910±0.004	0.633±0.051	0.913±0.004	0.612±0.073	0.912±0.002	0.606±0.078	0.913±0.003	0.639±0.069
ACI-Simple-0.005	0.975±0.003	0.207±0.050	0.977±0.003	0.194±0.035	0.978±0.003	0.185±0.041	0.972±0.006	0.269±0.088
ACI-Momentum-0.005	0.976±0.005	0.206±0.063	0.980±0.004	0.153±0.043	0.980±0.002	0.167±0.052	0.977±0.006	0.227±0.099
ACI-Simple-0.01	0.986±0.003	0.131±0.029	0.987±0.003	0.114±0.041	0.988±0.002	0.115±0.034	0.985±0.005	0.161±0.080
ACI-Momentum-0.01	0.987±0.004	0.121±0.065	0.990±0.002	0.088±0.030	0.989±0.003	0.102±0.044	0.993±0.003	0.069±0.026
CoFact	0.895±0.026	0.715±0.031	0.897±0.022	0.718±0.030	0.894±0.018	0.715±0.031	0.900±0.019	0.714±0.036

Table 7: Comparison with the method handling online conformal prediction on WikiData dataset.

	Lin		Squ		Sin		Ber	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
ACI-Simple-0.0005	0.894±0.004	0.766±0.004	0.895±0.005	0.765±0.005	0.894±0.004	0.764±0.003	0.892±0.006	0.760±0.011
ACI-Momentum-0.0005	0.894±0.004	0.765±0.004	0.896±0.004	0.765±0.005	0.895±0.004	0.764±0.004	0.893±0.006	0.760±0.012
ACI-Simple-0.001	0.901±0.003	0.755±0.007	0.902±0.004	0.751±0.006	0.902±0.003	0.750±0.006	0.903±0.005	0.746±0.015
ACI-Momentum-0.001	0.902±0.002	0.753±0.010	0.903±0.003	0.750±0.009	0.903±0.002	0.750±0.007	0.903±0.005	0.745±0.017
ACI-Simple-0.005	0.940±0.006	0.603±0.081	0.940±0.007	0.612±0.079	0.941±0.008	0.596±0.097	0.945±0.015	0.533±0.134
ACI-Momentum-0.005	0.944±0.007	0.586±0.080	0.945±0.006	0.578±0.089	0.944±0.007	0.577±0.101	0.948±0.016	0.508±0.148
ACI-Simple-0.01	0.952±0.011	0.504±0.117	0.952±0.011	0.509±0.123	0.951±0.013	0.514±0.127	0.961±0.011	0.440±0.124
ACI-Momentum-0.01	0.966±0.012	0.387±0.140	0.968±0.005	0.339±0.084	0.968±0.004	0.335±0.078	0.971±0.016	0.297±0.116
CoFact	0.896±0.010	0.748±0.006	0.895±0.009	0.748±0.006	0.895±0.008	0.748±0.006	0.897±0.008	0.749±0.006

Table 8: Comparison with method handling robust online conformal prediction on MedLFQA dataset.

	Lin		Squ		Sin		Ber	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
rACI-0.05	0.893±0.002	0.343±0.081	0.894±0.002	0.322±0.078	0.894±0.002	0.321±0.057	0.894±0.001	0.358±0.058
rACI-0.10	0.891±0.004	0.285±0.075	0.891±0.003	0.294±0.059	0.891±0.003	0.298±0.061	0.892±0.003	0.273±0.064
rACI-0.15	0.888±0.006	0.271±0.071	0.887±0.007	0.263±0.079	0.887±0.006	0.271±0.065	0.889±0.005	0.248±0.070
rACI-0.20	0.884±0.009	0.249±0.080	0.886±0.005	0.233±0.050	0.886±0.006	0.236±0.057	0.886±0.006	0.226±0.062
CoFact	0.895±0.026	0.715±0.031	0.897±0.022	0.718±0.030	0.894±0.018	0.715±0.031	0.900±0.019	0.714±0.036

Table 9: Comparison with method handling robust online conformal prediction on WikiData dataset.

	Lin		Squ		Sin		Ber	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
rACI-0.05	0.895±0.001	0.731±0.021	0.895±0.001	0.737±0.018	0.895±0.001	0.732±0.019	0.895±0.000	0.738±0.015
rACI-0.10	0.891±0.001	0.741±0.016	0.890±0.001	0.742±0.016	0.891±0.001	0.740±0.016	0.890±0.001	0.728±0.021
rACI-0.15	0.886±0.001	0.717±0.027	0.886±0.001	0.717±0.028	0.886±0.001	0.719±0.030	0.886±0.002	0.713±0.029
rACI-0.20	0.880±0.001	0.719±0.025	0.881±0.001	0.712±0.027	0.881±0.001	0.715±0.022	0.879±0.002	0.734±0.015
CoFact	0.896±0.010	0.748±0.006	0.895±0.009	0.748±0.006	0.895±0.008	0.748±0.006	0.897±0.008	0.749±0.006

From the results, we can draw the following observations. First, the ability of ACI to precisely control the factuality level is highly sensitive to the choice of the step size γ. A step size that is too small results in poor control of factuality, while a step size that is too large produces overly conservative prediction sets with a low number of retained claims. Although the original paper recommends setting $\gamma = 0.005$, our results indicate that this choice leads to overly conservative filtering in our scenario. This suggests that ACI requires extensive tuning of γ to achieve reasonable performance across different contexts. Secondly, although CoFact does not rely on access to true labels during the online learning process, it still achieves comparable factuality control performance to ACI with well-tuned γ values, further underscoring its effectiveness in our online setting. It is important to note that while ACI can achieve controlled factuality and, in some cases, retain slightly more claims than CoFact, this does not imply that ACI is superior. ACI relies on access to true labels after making predictions at each time step, a requirement that is impractical in our online setting.

F.3 Comparison with Robust Online Conformal Prediction

In this section, we relax the assumption of inaccessible true labels and instead assume access to noisy labels in the online conformal prediction setting. Under this relaxed setting, we compare our method with rACI, an extension of ACI proposed by Xi et al. (2025), which is designed to handle label noise.

To simulate a weak supervision scenario, we introduced uniform random noise to the true factuality labels in our experiments and compared the performance of CoFact with rACI. The results are presented in Table 8 and Table 9 for the MedLFQA and WikiData datasets, respectively. The noise ratio was varied from 5% to 20% to evaluate the robustness of both methods under different noise conditions.

Table 10: Experiments with varying training fractions on MedLFQA dataset.

	Lin		**Squ**		**Sin**		**Ber**	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
train_frac=0.05	0.920±0.028	0.613±0.094	0.921±0.024	0.612±0.097	0.923±0.023	0.622±0.099	0.913±0.049	0.621±0.189
train_frac=0.1	0.902±0.015	0.669±0.050	0.900±0.019	0.674±0.053	0.904±0.018	0.683±0.050	0.911±0.029	0.673±0.126
train_frac=0.3	0.890±0.014	0.722±0.057	0.889±0.016	0.721±0.056	0.891±0.018	0.733±0.057	0.911±0.015	0.666±0.047

Table 11: Experiments with varying training fractions on WikiData dataset.

	Lin		**Squ**		**Sin**		**Ber**	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
train_frac=0.05	0.899±0.012	0.744±0.019	0.898±0.012	0.744±0.019	0.898±0.011	0.743±0.020	0.891±0.013	0.760±0.016
train_frac=0.1	0.905±0.013	0.741±0.013	0.905±0.013	0.741±0.013	0.905±0.013	0.741±0.013	0.897±0.015	0.749±0.017
train_frac=0.3	0.892±0.012	0.751±0.011	0.892±0.013	0.750±0.011	0.893±0.013	0.750±0.011	0.894±0.010	0.754±0.009

Table 12: Experiments with varying expert number base on MedLFQA dataset.

	Lin		**Squ**		**Sin**		**Ber**	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
expert_num_base=3	0.883±0.016	0.716±0.022	0.888±0.019	0.712±0.020	0.887±0.022	0.714±0.022	0.896±0.012	0.713±0.032
expert_num_base=5	0.888±0.022	0.720±0.026	0.887±0.022	0.719±0.025	0.891±0.021	0.723±0.028	0.899±0.020	0.706±0.032
expert_num_base=10	0.892±0.020	0.715±0.025	0.896±0.020	0.714±0.024	0.900±0.021	0.714±0.022	0.894±0.014	0.713±0.029

Table 13: Experiments with varying expert number base on WikiData dataset.

	Lin		**Squ**		**Sin**		**Ber**	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
expert_num_base=3	0.900±0.011	0.745±0.009	0.899±0.011	0.744±0.008	0.899±0.011	0.745±0.009	0.893±0.015	0.750±0.009
expert_num_base=5	0.895±0.007	0.746±0.010	0.894±0.007	0.748±0.010	0.894±0.007	0.747±0.010	0.889±0.013	0.746±0.008
expert_num_base=10	0.901±0.016	0.745±0.008	0.903±0.015	0.746±0.007	0.903±0.015	0.746±0.008	0.901±0.012	0.745±0.005

From these results, we observe that while rACI exhibits some robustness to label noise, its performance deteriorates significantly as the noise level increases. Notably, on the MedQA dataset, the Claims Retained metric for rACI is extremely low, making it impractical for real-world applications. Furthermore, rACI relies on strong assumptions about the nature of noise, specifically that label-flipping probabilities are consistent across different sub-claims and samples, and that the true noise rates are known. These assumptions are rarely realistic in practical scenarios. Given these limitations, CoFact demonstrates clear advantages in terms of practicality and robustness for factuality control in online settings.

F.4 Ablation Study on Calibration Sample Size

We conducted experiments using different calibration set ratios, specifically $\{0.05, 0.1, 0.2\}$, on the MedLFQA dataset (corresponding to approximately 240, 480, and 1440 samples) and the WikiData dataset (approximately 425, 850, and 2550 samples) under four shift types. These experiments were designed to assess the impact of calibration set size on the performance of our method. The results for the two datasets are summarized in Table 10 and Table 11, respectively.

The results indicate that our method exhibits strong robustness to variations in calibration set size. Even with a relatively small calibration set (approximately 240 samples for MedQA and 425 samples for Wiki), our method achieves satisfactory performance in terms of both Factuality and Claims Retained. This underscores the practicality of our approach, particularly in scenarios where only limited calibration data is available.

F.5 Ablation Study on Varying Expert Number

To investigate the impact of the number of experts, we adjusted the expert lifetime schedule by modifying the lifetime scaling factor from 2^i to 3^i, 5^i, and 10^i. This change alters the number of active experts at each time step to $\log_3 t$, $\log_5 t$, and $\log_{10} t$, respectively. The experimental results on the MedLFQA and WikiData datasets, evaluated under four shift types, are summarized in Table 12 and Table 13.

The results show that our method consistently achieves reliable performance in terms of both Factuality and Claims Retained across different expert base settings. This indicates that the method is robust to variations in the number of active experts, further highlighting its adaptability to different configurations.

Table 14: Experiments with varying steps per update on MedLFQA dataset.

	Lin		Squ		Sin		Ber	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
per_step_num=5	0.893±0.024	0.723±0.020	0.890±0.025	0.722±0.027	0.887±0.029	0.726±0.023	0.896±0.014	0.712±0.030
per_step_num=10	0.897±0.024	0.703±0.028	0.895±0.025	0.701±0.028	0.897±0.025	0.709±0.027	0.902±0.012	0.696±0.027
per_step_num=20	0.899±0.024	0.696±0.031	0.900±0.025	0.693±0.032	0.899±0.025	0.698±0.030	0.904±0.011	0.684±0.028

Table 15: Experiments with varying steps per update on WikiData dataset.

	Lin		Squ		Sin		Ber	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
per_step_num=5	0.900±0.011	0.750±0.008	0.899±0.010	0.750±0.010	0.901±0.011	0.747±0.008	0.893±0.016	0.749±0.011
per_step_num=10	0.903±0.009	0.748±0.006	0.905±0.009	0.749±0.006	0.901±0.008	0.748±0.007	0.897±0.008	0.747±0.011
per_step_num=20	0.899±0.007	0.748±0.006	0.901±0.006	0.747±0.006	0.900±0.007	0.748±0.007	0.900±0.013	0.751±0.011

Table 16: Experiments with varying noise levels on MedLFQA Dataset.

	Lin		Squ		Sin		Ber	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
noise_level=0.05	0.898±0.021	0.722±0.030	0.897±0.018	0.720±0.028	0.899±0.021	0.725±0.027	0.909±0.016	0.713±0.027
noise_level=0.10	0.892±0.024	0.743±0.041	0.891±0.021	0.742±0.040	0.893±0.025	0.746±0.039	0.902±0.021	0.734±0.035
noise_level=0.15	0.884±0.028	0.776±0.048	0.879±0.024	0.776±0.047	0.882±0.028	0.781±0.046	0.893±0.026	0.768±0.041

Table 17: Experiments with varying noise levels on the WikiData Dataset.

	Lin		Squ		Sin		Ber	
	Factuality	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**	**Factuality**	**Claims Retained**
noise_level=0.05	0.898±0.006	0.755±0.008	0.897±0.005	0.755±0.008	0.898±0.005	0.755±0.007	0.898±0.003	0.752±0.006
noise_level=0.10	0.894±0.008	0.761±0.010	0.893±0.008	0.761±0.010	0.894±0.008	0.761±0.009	0.895±0.009	0.758±0.009
noise_level=0.15	0.892±0.009	0.765±0.010	0.891±0.009	0.765±0.010	0.892±0.008	0.765±0.010	0.894±0.010	0.761±0.011

F.6 Ablation Study on Varying steps per update

To further evaluate the applicability of our method in an online batch learning setup, we performed experiments using different batch sizes ($\{1, 5, 10, 20\}$) on the MedLFQA and WikiData datasets across four shift types. The results are summarized in Table 14 and Table 15.

The findings indicate that our method consistently maintains strong performance in terms of both Factuality and Claims Retained when processing incoming samples in batches. This demonstrates the suitability of our approach for online batch learning scenarios.

F.7 Experiments on Varying Noise Levels in Feature Representation

We conducted experiments to evaluate the sensitivity of our method to feature representations by introducing Gaussian noise with varying levels of variance to the feature representations on the MedLFQA and WikiData datasets. The results are summarized in Table 16 and Table 17.

To ensure a meaningful analysis, the feature representations were normalized prior to adding noise, resulting in a median feature value of approximately 0.15. As such, introducing a noise level of 0.15 constitutes a substantial perturbation to the features.

The results show that even under relatively high noise levels, our method consistently maintains robust performance in terms of both Factuality and Claims Retained. This highlights the resilience of our approach to variations and noise in feature representations.

G More Case Studies

Prompt: *What is MVC in software engineering?*

Filtered Response: MVC stands for Model-View-Controller, which is a design pattern commonly used in software engineering.

- Model: Represents the data and behavior of the application. ~~It acts as a bridge between the View and the Controller~~, handling the business logic and data manipulation.

- View: Represents the user interface of the application. It displays the data to the user and sends user inputs to the Controller.

- Controller: Acts as an intermediary between the Model and the View. It processes user input from the View, interacts with the Model to retrieve or update data, ~~and updates the View accordingly~~.

MVC helps to separate concerns and facilitate maintainability, reusability, and scalability of the codebase. ~~It is commonly used in web development frameworks like Ruby on Rails, ASP.NET MVC, and Spring MVC.~~

In this case, CoFact effectively removes the hallucinated claims ”It acts as a bridge between the View and the Controller” while retaining most of the correct claims.

H Limitations and Future Work

CoFact’s theoretical guarantees apply to the entire time horizon and may not hold for smaller time intervals. Developing methods that provide finer-grained factuality guarantees is an important direction for future work. Additionally, while CoFact focuses on ensuring the factuality of filtered claims, other response qualities, such as informativeness and diversity, could also be required in certain scenarios. Extending CoFact to incorporate these aspects presents another promising avenue for future research.