Audit Me If You Can: Query-Efficient Active Fairness Auditing of Black-Box LLMs

David Hartmann[1,2], Lena Pohlmann[1,2], Lelia Hanslik[2], Noah Gießing[3], Bettina Berendt[1,2,4], Pieter Delobelle[0,4,5]
[1]Weizenbaum Institut Berlin, [2]Technische Universität Berlin, [3]FIZ Karlsruhe, [4]KU Leuven [5]Pleias

Abstract

Large Language Models (LLMs) exhibit systematic biases across demographic groups. Auditing is proposed as an accountability tool for black-box LLM applications, but suffers from resource-intensive query access. We conceptualise auditing as uncertainty estimation over a target fairness metric and introduce BAFA, the Bounded Active Fairness Auditor for query-efficient auditing of black-box LLMs. BAFA maintains a version space of surrogate models consistent with queried scores and computes uncertainty intervals for fairness metrics (e.g., Δ_{AUC}) via constrained empirical risk minimisation. Active query selection narrows these intervals to reduce estimation error. We evaluate BAFA on two standard fairness dataset case studies: CIVILCOMMENTS and BIAS-IN-BIOS, comparing against stratified sampling, power sampling, and ablations. BAFA achieves target error thresholds with up to 41× fewer queries than stratified sampling (e.g., 144 vs 5,956 queries at $\varepsilon = 0.02$ for CIVILCOMMENTS) for tight thresholds, demonstrates substantially better performance over time, and shows lower variance across runs. These results suggest that active sampling can reduce resources needed for independent fairness auditing with LLMs, supporting continuous model evaluations.

1 Introduction

LLMs are increasingly deployed not only for generative tasks such as text completion, image synthesis, and video generation, but also for downstream decision-making tasks, including classification, scoring, and ranking. These systems are commonly offered via machine-learning-as-a-service (MLaaS) APIs and have substantial real-world impact, for example, in automated hate speech detection and candidate screening in hiring.

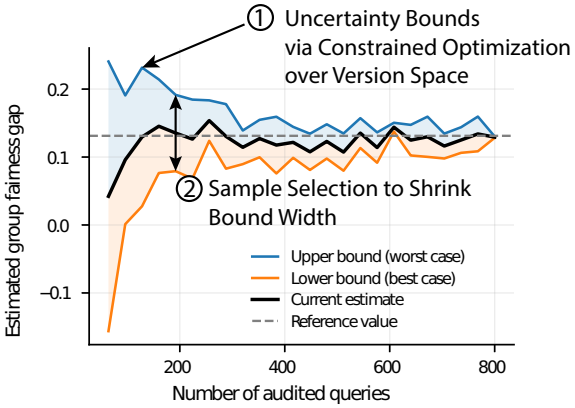

Figure 1: **Bounded Active Fairness Auditing (BAFA)**. Upper and lower bounds on the fairness metric converge as queries accumulate. BAFA aims to maximally shrink the uncertainty interval between bounds.

However, recent evaluations have shown that such applications exhibit systematic performance disparities across social groups. Commercial hate speech detection systems based on black-box LLMs have been found to underperform for LGBTQIA+ and people with disabilities (Röttger et al., 2021; Hartmann et al., 2025b). Similarly, LLM-based CV and biography screening systems show biases with respect to disability status (Glazko et al., 2024), gender (Wang et al., 2024), or educational background (Iso et al., 2025).

To uncover such systemic risks in deployed systems, audits have been proposed as a key accountability mechanism (Raji et al., 2020; Birhane et al., 2024). Independent black-box auditing is increasingly reflected in policy frameworks, including Appendix 3.5 of the EU Code of Practice on Generative AI, and it is widely discussed in governance and regulatory proposals (Mökander et al., 2024; Raji et al., 2022; Hartmann et al., 2025a).

In practice, however, conducting fairness audits of black-box LLMs remains challenging. Compre-

[0]Work done while at Aleph Alpha

Findings of the Association for Computational Linguistics: ACL 2026, pages 33673–33698
July 2-7, 2026

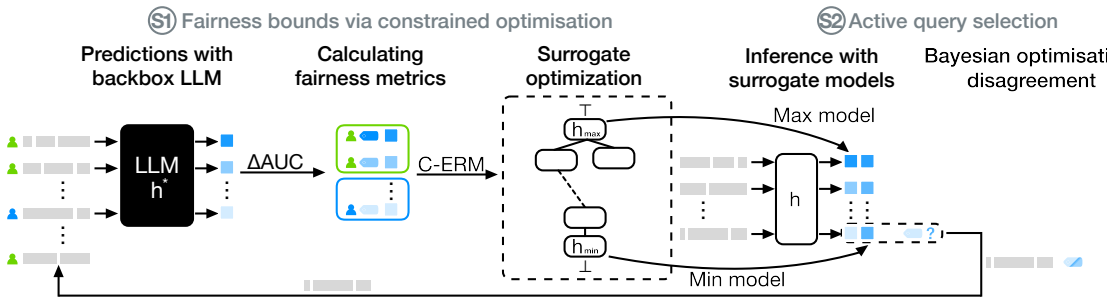

Figure 2: **BAFA Pipeline in more detail**. In every turn we sample k samples. First, we query the black-box LLM with a stratified seed set from our dataset (1). Then, we calculate the estimated fairness measure (2) and do constraint optimisation with BERT surrogates (3) to get lower and upper fairness bounds. Based on the calculated scores for each $x \in D$ from the upper and lower models (4), BAFA selects queries (5) which shrink the distance between the lower and upper in high-disagreement regions, leading to faster and more stable convergence.

hensive audits typically require extensive amount of API queries, which are costly (Hartmann et al., 2025b), may raise privacy concerns (Zaccour et al., 2025), and can conflict with data minimisation obligations under the GDPR (Rastegarpanah et al., 2021).

These challenges are especially pronounced in continuous auditing scenarios, where linguistic change and system updates necessitate repeated evaluations over time. A common workaround is the use of hand-crafted bias benchmarks or templates (*e.g.,* Röttger et al., 2021; Nadeem et al., 2021; Gehman et al., 2020; Nangia et al., 2020; Röttger et al., 2022). While useful for controlled testing, these approaches often lack ecological validity and provide limited reliability when auditing dynamic, context-sensitive tasks such as hate speech detection or biography scoring in real-world settings (Delobelle et al., 2024).

Several works therefore argue for query-efficient, ecologically valid fairness auditing that operates under strict budget and black-box access constraints, enabling continuous evaluation by independent auditors (Cen and Alur, 2024; Hartmann et al., 2025b). While Yan and Zhang (2022) propose an oracle-efficient method for auditing demographic parity, their approach does not scale to large hypothesis spaces such as LLMs and is incompatible with ranking-based fairness metrics commonly used in content moderation and hiring. Related work on query-efficient red teaming (*e.g.,* Lee et al., 2023) actively surfaces harmful behaviours but cannot estimate specific fairness parameters in black-box settings. This gap motivates a query-efficient active fairness auditing method for black-box LLMs that supports ranking metrics.

Our approach. Bounded Active Fairness Auditing is introduced as a query-efficient auditing framework for black-box language models. As illustrated in Figure 1, BAFA conceptualises auditing as measuring uncertainty associated with a model's group fairness parameter, such as the group-wise ROC-AUC difference, within a specified query budget. This is achieved by optimising surrogate upper and lower bounds, colour-marked in one illustrative run. These represent an interval of plausible values for a fairness metric, based on the outputs obtained thus far.

BAFA then actively selects queries that are expected to most effectively reduce uncertainty regarding the target fairness metric, thereby minimising the required query budget as demonstrated in the BAFA pipeline in Figure 2. By focusing queries on fairness-critical regions of the input space, BAFA significantly reduces audit costs. Experimental results show that BAFA requires substantially fewer queries than baselines, performing better over time and achieving lower variance, evaluated in two practical auditing scenarios.

The contributions of this work are threefold: (1) **methodological:** we present an active fairness auditing method for black-box LLMs that works with threshold-invariant ranking metrics – to the best of our knowledge, the first active learning approach for black-box LLM fairness auditing, (2) **practical:** we introduce a query-efficient framework suitable for independent audits under limited access, budget, and regulatory constraints, and (3) **empirical:** we compare several samplings methods for auditing and substantial query-efficiency gains over three baseline sampling approaches in two realistic LLM auditing case studies.

2 Related Work

Fairness evaluation of language models. LLMs exhibit systematic demographic biases (Blodgett et al., 2020), as demonstrated by hate speech detection (Sap et al., 2019) and CV scoring (Glazko et al., 2024). Prior work has relied heavily on template-based benchmarks such as HateCheck (Röttger et al., 2021), StereoSet (Nadeem et al., 2021), and RealToxicityPrompts (Gehman et al., 2020), which enable controlled comparisons but suffer from limited construct validity and weak alignment with real-world use (Goldfarb-Tarrant et al., 2021). As a result, these benchmarks provide static snapshot evaluations that are ill-suited for auditing deployed systems over time (Tonneau et al., 2025). Critically, Blodgett et al. (2021) argue that benchmark-driven evaluations often conflate distinct notions of bias and obscure concrete group-level harms, motivating auditing approaches grounded in real-world data and explicit fairness metrics.

Black-box auditing and red teaming. Under black-box access and beyond benchmark-driven evaluation, two dominant evaluation paradigms have emerged: red teaming and auditing. Red teaming seeks to uncover worst-case or unsafe behaviours through adversarial querying, providing evidence of failure modes without estimating their prevalence (Perez et al., 2022). In contrast, black-box auditing aims to estimate well-defined system properties, such as fairness, via systematic black-box queries, potentially conducted by independent external stakeholders (Raji et al., 2022; Mökander et al., 2024). However, comprehensive audits are often infeasible in practice due to high query costs, rate limits, and legal constraints such as GDPR data minimisation (Rastegarpanah et al., 2021; Zaccour et al., 2025). These constraints motivate query-efficient auditing methods that can provide reliable estimates within strict budgets.

Query-efficient and active auditing. Query-efficient auditing seeks to estimate a fairness measure of a black-box using as few queries as possible. Existing work has focused on sample size reduction via rigorous passive sampling approaches. For example, Singh et al. (2023) derive closed-form requirements for detecting fairness violations under power sampling, but do not consider adaptive query selection. While active learning reduces label complexity by selecting informative examples (Settles, 2009), most approaches optimise predictive performance rather than group-level fairness estimation. Recent frameworks cover related areas, like online monitoring with confidence sequences (Maneriker et al., 2023), Fourier fairness coefficients for discretised inputs, Ajarra et al. (2024), and Bayesian Optimisation (BO) for red teaming (Lee et al., 2023). However, they do not support group fairness auditing when statistical uncertainties are present. Active fairness auditing using constrained empirical risk minimisation (C-ERM) (Yan and Zhang, 2022) offers strong guarantees for threshold-based metrics. However, it depends on optimisation surrogates that are not practical for modern LLMs, since these surrogates must closely mimic the black-box model. Most active auditing frameworks also focus on threshold-dependent classification metrics, even though many commercial models produce continuous scores. Both Yan and Zhang (2022) and Singh et al. (2023) have called for extensions to these metrics. For such systems, threshold-invariant measures like group-wise ROC AUC difference are more suitable, as they capture disparities across all possible decision thresholds (Borkan et al., 2019a,b; Gallegos et al., 2024).

3 Bounded Active Fairness Auditing

Black-box Audit Setup. We audit a black-box model $h^\star$ that assigns scores to inputs (e.g., toxicity scores for comments, confidence scores for occupation predictions). Given labeled data with ground-truth labels y_i and protected group attributes $g_i \in \{0, 1\}$, our goal is to estimate the ranking fairness gap between two demographic groups:

$$\Delta_{\text{AUC}}(h^\star) = \text{AUC}_{g=0}(h^\star) - \text{AUC}_{g=1}(h^\star),$$

where AUC_g measures how well the model ranks positive examples above negative examples for group g. Given a query budget T (e.g., 1000 API calls), we seek an estimator $\widehat{\Delta}_{\text{AUC}}$ that is ϵ-accurate (e.g., within ± 0.02 of the true disparity) while minimising the number of queries $q \leq T$ needed. We assume access to ground-truth labels and group attributes for evaluation, but only black-box access to the model itself (complete mathematical formulation can be found in App. A.1).

Algorithm Overview. Figure 2 summarises our method, Bounded Active Fairness Auditing (BAFA)[1]. Starting from a stratified seed set, BAFA

[1] https://github.com/dawiethart/AuditMeIfYouCan-Active-Auditing

Algorithm 1 Bounded Active Fairness Auditing (BAFA)

Require: Audit pool $\mathcal{U}$, black-box $h^\star$, budget T, tolerance λ, batch size k, threshold ϵ, fairness measure μ, hypothesis space $\mathcal{H}$

$S_0 \leftarrow$ stratified seed set; query $h^\star$ to attach scores $\{s_i^\star\}$
for $t = 0, 1, 2, \ldots, T$ **do**
  // S1: Certificate
  Solve $h^t_{\max} \leftarrow \arg\max_{h \in \mathcal{H}_\lambda(S_t)} \mu(h)$
  Solve $h^t_{\min} \leftarrow \arg\min_{h \in \mathcal{H}_\lambda(S_t)} \mu(h)$
  $[\mu^t_{\min}, \mu^t_{\max}] \leftarrow [\mu(h^t_{\min}),\ \mu(h^t_{\max})]$
  if $(\mu^t_{\max} - \mu^t_{\min})/2 \le \epsilon$ **then**
    return $\hat{\mu}_t = (\mu^t_{\min} + \mu^t_{\max})/2$
  end if
  // S2: Active query selection
  for each $x \in \mathcal{U} \setminus S_t$ **do**
    $\mathrm{dis}_t(x) \leftarrow |h^t_{\max}(x) - h^t_{\min}(x)|$
  end for
  $Q_t \leftarrow$ top-k candidates by $\mathrm{dis}_t(x)$
  Query $h^\star$ on Q_t; update $S_{t+1} \leftarrow S_t \cup Q_t$
end for
// Implementation details: App. A.2

iteratively (**S1**) computes upper and lower fairness bounds via constrained optimisation, and (**S2**) selects new queries that are expected to maximally reduce the bound width and thus, uncertainty in the group fairness measure.

S1: Fairness bounds via constrained optimisation. BAFA quantifies uncertainty in fairness by maintaining a set of surrogate hypotheses that are consistent with the black-box model on the queried set S. Specifically, BAFA maintains a λ-approximate version space $\mathcal{H}_\lambda(S_t) = \{h \in \mathcal{H} : |h(x_i) - s_i^\star| \le \lambda,\ \forall (x_i, s_i^\star) \in S_t\}$ of BERT-based surrogates (Devlin et al., 2019) consistent with observed black-box scores. In each round, we solve two constrained problems via Cooper (Gallego-Posada et al., 2025) where one is maximising, and one is minimising Δ_{AUC} over $\mathcal{H}_\lambda(S_t)$. This yields extremal hypotheses $h^t_{\max}, h^t_{\min}$ and the certificate interval $[\mu^t_{\min}, \mu^t_{\max}]$. As ROC-AUC is non-differentiable, we optimise a sigmoid pairwise ranking surrogate (Agarwal, 2013). As $|S_t|$ grows, $\mathcal{H}_\lambda(S_t)$ shrinks monotonically and the interval narrows (see App. A.1).

S2: Active query selection. To reduce required queries, BAFA selects inputs expected to shrink the fairness uncertainty interval the most. We operationalise this by scoring each unqueried candidate via the disagreement of the two extremal surrogates from S1:

$$\mathrm{dis}_t(x) = \left|h^t_{\max}(x) - h^t_{\min}(x)\right|,$$

and querying the top-k highest-scoring inputs per round. We adopt top-k rather than the ε-driven disagreement loop of Yan and Zhang (2022), which iterates until the surrogate itself falls below a diameter threshold. Such schemes require the surrogate to reliably mimic the black box. This condition holds only after roughly 500–750 queries in our setting (App. A.6.2). Top-k instead uses the surrogate *only* to rank candidates by bound disagreement, not as a black-box proxy. This is a strictly weaker requirement as it demands consistent *ordering* of candidates rather than accurate *score replication*, and remains effective even under partial surrogate–black-box agreement of scores (App. A.6.2).

Two disagreement-based scoring rules are used and evaluated that do not require an accurate surrogate for query selection. First, *Bound-disagreement sampling* prioritises candidates where the current upper- and lower-bound models of S1 disagree most on AUC-relevant pairwise rankings. Second, as a comparison to *Bound-disagreement sampling*, we test *Bayesian optimisation* which searches over acquisition features – including bound disagreement, LoRA-surrogate diversity, and surrogate–black-box disagreement – as a proxy for uncertainty. Inspired by Lee et al. (2023), this should balance between exploitation of high-impact regions and exploration for text diversity. For both strategies, we apply distributional regularisation using empirical subgroup and label marginals to mitigate selection-induced bias in fairness estimation (Details, see App. A.2).

4 Experimental Setup

BAFA is evaluated in two black-box LLM deployments under realistic audit constraints: (A) hate speech detection and (B) profession estimation from biographies. The two case studies are deliberately complementary: Case Study A uses a locally-hosted classifier with a known, synthetically injected disparity ($\mu_{\Delta\mathrm{AUC}} \approx 0.14$) and an audit

Case Study	ε	BAFA (Dis.)	BAFA (BO)	C-ERM (abl.)	BO only (abl.)	Power (base.)	Stratified (base.)
Queries to ε (mean [95% CI]) ↓							
CIVILCOMMENTS	0.02	**144 [98, 190]**	256 [180, 332]	457 [320, 594]	1,204 [864, 1,544]	8,436 [7,876, 8,836]	5,956 [1,268, 7,652]
	0.05	**80 [58, 102]**	132 [96, 168]	137 [102, 172]	356 [240, 472]	932 [484, 2,756]	452 [164, 676]
BIAS-IN-BIOS	0.02	**340 [248, 432]**	356 [268, 444]	512 [380, 644]	772 [580, 964]	5,396 [516, 5,988]	1,748 [564, 3,060]
	0.05	148 [108, 188]	180 [136, 224]	210 [158, 262]	**100 [72, 128]**	356 [4, 372]	212 [4, 324]
Mean AUEC for first 1k queries ↓							
CIVILCOMMENTS		**0.019**	0.022	0.030	0.060	0.093	0.066
BIAS-IN-BIOS		**0.025**	0.029	0.042	0.035	0.045	0.042
Error at 250 queries (mean [95% CI]) ↓							
CIVILCOMMENTS		**0.020 [0.015, 0.026]**	0.021 [0.013, 0.030]	0.030 [0.015, 0.046]	0.096 [0.062, 0.131]	0.108 [0.080, 0.135]	0.064 [0.046, 0.083]
BIAS-IN-BIOS		0.022 [0.017, 0.027]	**0.022 [0.017, 0.026]**	0.024 [0.014, 0.033]	0.023 [0.014, 0.033]	0.065 [0.045, 0.085]	0.043 [0.028, 0.058]

Table 1: **BAFA substantially reduces query costs in both case studies while beating baselines in over-time performance and stability across 20 seeds.** We report (i) *convergence query-efficiency* as queries required until the mean error curve falls under ε, with bootstrapped 95% CIs across 20 seeds; (ii) *over-time performance* via AUEC over the first 1k queries; and (iii) *mid-budget error* at 250 queries with bootstrapped 95% CIs. Dis. = disagreement-based selection. Wide CIs for baselines at strict thresholds reflect high run-to-run variability, further supporting the case for active auditing.

pool of ∼50k comments, providing a controlled setting with a severe fairness violation; Case Study B uses a commercial black-box (GPT-4.1-mini) with a non-injected disparity ($\mu_{\Delta\mathrm{AUC}} \approx 0.02$–$0.045$) and scores over ∼50k biographies, testing BAFA under architectural mismatch between surrogate and target. Together, they span the range of realistic independent audit settings with two high-stakes scenarios that should be independently evaluated.

In both case studies, the auditor has access only to model inputs and outputs and seeks to estimate group-level ROC AUC disparities under a fixed query budget. All strategies are evaluated using a common protocol with identical budgets and batch sizes, and the results are averaged across 20 random seeds. At each audit round, a batch of inputs is selected for black-box querying with each strategy, and the fairness estimate is updated.

In Table 1, we report *convergence query-efficiency* as the number of black-box queries needed until the *mean* absolute error across seeds first falls below a target threshold $\varepsilon \in \{0.02, 0.05\}$. Additionally, we report *over-time performance* via the area under the error curve (AUEC) over the first 1000 queries (analogous to AUC), and quantify *stability* by the mean error and standard deviation across seeds at fixed budgets. We compare BAFA against stratified and power sampling (calculated for Δ_{AUC} from Singh et al. (2023)) as baselines, constrained optimisation (as in Yan and Zhang (2022) with a stratified sample), and BO without active querying as ablations. These baselines and ablations allow us to disentangle the effects of active selection and constrained optimisation. Complete evaluation metric details (App. A.5.1), baseline and ablations definitions (App. A.3.1) as well as implementation details (App. A.4) are provided in Appendix. A sensitivity analysis covering λ, batch size k, regularisation weight α, and C-ERM optimization epochs is provided in App. A.4 (Tables 5–6).

5 Results

Table 1 summarises the query efficiency and estimation performance of different auditing strategies over 20 random seeds in both case studies.

5.1 Case Study A: Auditing Hate Speech Detection

Our first case study audits group-based performance disparities in hate speech detection using real-world, identity-labelled data. We use the CIVILCOMMENTS dataset (Borkan et al., 2019b), which contains user-generated public comments on English-language news sites, annotated for toxicity and multiple identity targets. We focus on eight target groups commonly studied in prior work (e.g., gender, religion, sexual orientation) and evaluate disparities between dominant and marginalised groups (For details see App. A.4).

As the audited system, we construct a controlled but highly biased black-box model by fine-tuning HateBERT (Caselli et al., 2021) on the SBIC dataset (Sap et al., 2020), systematically flipping labels for comments targeting marginalised groups ($\mu_{\Delta AUC} \approx 0.14$ for each group pair). This synthetic setup provides a known and severe fairness violation, allowing us to assess whether active au-

diting can reliably detect disparities under limited query budgets.

Query efficiency to target threshold. Across both thresholds, active auditing strategies require substantially fewer queries than passive baselines to reach a given accuracy. For $\varepsilon = 0.02$, BAFA with disagreement and BAFA with BO reach the target error within 144–256 queries on average, whereas stratified and power sampling require several thousand queries, ablations around 2–8$\times$ more. Disagreement-based sampling is approximately 41$\times$ faster than stratified sampling for $\varepsilon = 0.02$. The result is a bit less pronounced for the less stricter threshold $\varepsilon = 0.05$, where both BAFA approaches reduce the mean of queries needed around three to five times (5.7 for disagreement and 3.4 for BO) in relation to stratified sampling.

Over-time estimation accuracy. Presented in Figure 3 and by mean AUEC, our active methods also perform substantially better (error-reduction around 3-4 times) than baselines in terms of over-time performance. Over the first 1,000 queries, BAFA with disagreement achieves the lowest mean AUEC on CIVILCOMMENTS, followed closely by BAFA with BO. In contrast, stratified and power sampling accumulate substantially higher error over time due to slow early progress, whereas after 1,000 queries, AUEC is similar across all approaches. Interestingly, C-ERM already outperforms both baselines (see Figure 3) and BO without active sampling, but its performance is still below that of the BAFA variants.

Mid-budget accuracy and stability. At a mid-range budget of 250 queries, BAFA with disagreement achieves the lowest mean estimation error on CIVILCOMMENTS with reduced variance across seeds, which are also visible in CI-bands in Figure 3, making it the most reliable estimator at fixed budgets. BAFA with BO is close in mean error but exhibits higher variance at this point, representing a trade-off between early exploration and overall stability, although mean AUEC are comparable across both BAFA approaches. Baselines demonstrate substantially larger error bands at 250 queries and show large run-to-run variability.

5.2 Case Study B: Auditing Black-Box CV Scoring LMs

Our second case study examines fairness in automated hiring scenarios by auditing a black-box

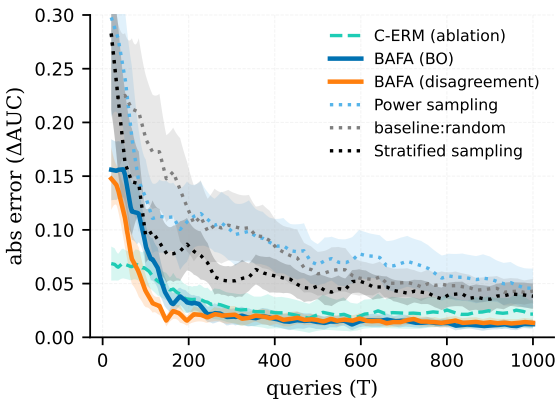

Figure 3: **Active auditing methods perform more query-efficient and stable over 20 CIVIL COMMENTS seeds.** BAFA methods (solid) converge significantly faster than baseline sampling strategies (dotted). Shaded areas indicate 95% confidence intervals across seeds and demonstrate that BAFA methods show substantially reduced variance compared to baseline methods.

language model used for occupation inference. We use the BIAS-IN-BIOS dataset (De-Arteaga et al., 2019), which contains short biographies annotated with ground-truth occupations and binary gender labels. We use GPT-4.1-mini as a black-box scorer via a deterministic prompt that maps biographies to (i) a predicted occupation from a predefined label set and (ii) a confidence score in $[0, 100]$.

A small, disjoint subset of biographies is used as few-shot examples to stabilise model behaviour; the remaining biographies form the audit dataset. For each occupation, we define a binary classification task (target occupation vs. all others), using the model's confidence score as a ranking signal. Group-wise ROC-AUCs are computed separately for male and female biographies, and fairness is again measured via Δ_{AUC} ($\mu_{\Delta AUC} \approx 0.02 - 0.045$) (Details App. A.4).

This case study complements content moderation by testing our method in a distinct, potentially biased domain with different data distributions and a commercial black-box model that is qualitatively different from our surrogate in both architecture and scale. One open question is whether BAFA performs better even for such substantially larger black-box models, since our C-ERM step uses a comparatively small BERT surrogate to reduce the *fairness-metric version space* induced by queried scores rather than the black box's full parameter space; we address this question empirically in this case study.

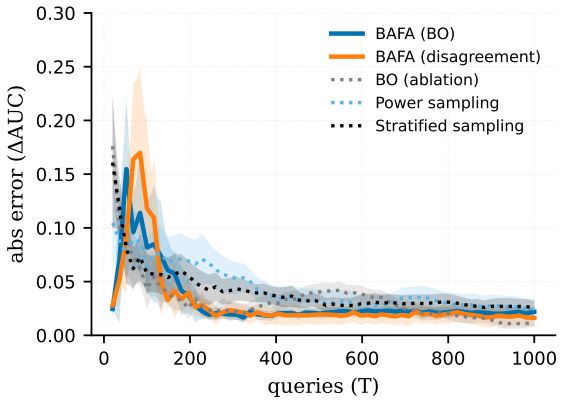

Figure 4: **Active auditing methods perform even with large parameter spaces with GPT-4.1-MINI as black-box.** Similarly, to Fig. 3, BAFA methods converge significantly faster than baseline sampling strategies and show substantially reduced variance compared to baseline methods. However, a much bigger variance and worse performance are visible for the first 100-120 queries, probably related to the model mismatch.

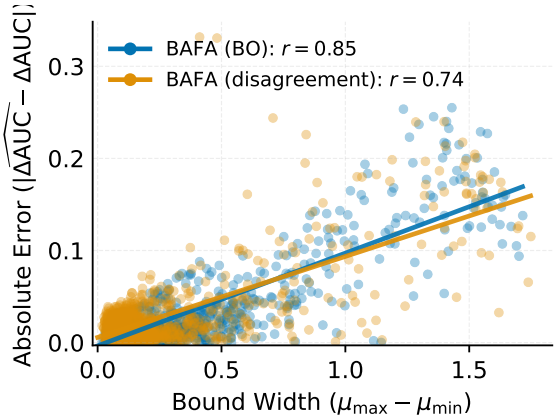

Figure 5: Relationship between BAFA's uncertainty bound width and the true absolute estimation error of ΔAUC in first 1k queries for CIVIL COMMENTS. Each point corresponds to an audit during active querying. Bound width strongly correlates with actual error for both BO and disagreement-based selection.

Query efficiency to target threshold. Again, on BIAS-IN-BIOS, active auditing converges faster than baselines when it comes to the strict accuracy thresholds. For $\varepsilon = 0.02$, both BAFA variants reach the target within around 340–356 queries on average, with disagreement performing slightly better than BO. Stratified sampling requires 1,748 queries and power sampling more than 5,300 queries, corresponding to roughly a $5\times$ and $16\times$ reduction, respectively. At the looser threshold $\varepsilon = 0.05$, BAFA's gains are smaller as it reaches the target within 148–180 queries, while baselines require 212 and 356 queries. Interestingly, for this threshold and case study, the ablation BO outperforms BAFA in convergence, although it needs about $2.3\times$ more queries for $\epsilon = 0.02$.

Over-time estimation accuracy and stability. Consistent with Case Study A, active methods achieve lower error throughout the audit process. BAFA with disagreement yields the lowest AUEC over the first 1,000 queries, indicating faster uncertainty reduction across rounds, while BAFA with BO performs comparably but with slightly higher cumulative error early on. However, we acknowledge that the difference to baselines is less pronounced than in case study A, and Figure 4 demonstrates that, although BAFA converges faster and is more stable after 100–120 queries, it shows more variance and larger error than baselines in the first 100–120 queries. At 250 queries, however, both BAFA variants achieve lower estimation error and higher stability than baselines.

6 Discussion

Active auditing can reduce query budget by a significant amount compared to baselines. Empirically, BAFA with both approaches reaches strict and loose targets with substantially fewer queries than baselines, maintains lower error throughout the audit, and is more stable at moderate budgets, especially on CIVILCOMMENTS, where baseline variability is high. Overall, the results indicate that BAFA is not just a query-efficient convergence algorithm, but a practical approach for producing more accurate and reproducible ΔAUC estimates under access and resource constraints. Furthermore, we observe a consistent trade-off between query selection methods: disagreement prioritises fast early interval shrinkage, while BO tends to achieve better mid-budget accuracy. Both approaches, however, seem to work similarly well, although our hypothesis was that BO would outperform simple disagreement. BAFA-BO, however, produces more reliable bound widths with high correlation to the absolute error.

Uncertainty calibration of BAFA. This view of auditing follows Yan and Zhang (2022) and treats uncertainty as a version space quantity by computing an uncertainty interval $[\mu_{\min}, \mu_{\max}]$ by solving two constrained optimisation problems that minimise and maximise the target metric over the ver-

sion space. The resulting width has a direct operational meaning as it upper-bounds how much the estimated $\Delta\mathrm{AUC}$ could change under any hypothesis still compatible with the observed queries, and it shrinks as additional queries eliminate hypotheses from the version space. While a fully Bayesian approach would instead report credible intervals from a posterior, the version space formulation is computationally tractable for auditing large black-box LLMs with small surrogates. Empirically, Figure 5 shows that bound width is strongly correlated with the true absolute estimation error, and that the ground-truth metric lies within BAFA's interval in over 95% of queries on CIVIL COMMENTS (99.9% for BAFA-BO and 95.4% for disagreement). On Bias-in-Bios, coverage is lower due to surrogate–target mismatch, but the average bound violation remains below our strict tolerance $\varepsilon = 0.02$, so width remains a useful proxy for uncertainty about the group-level fairness metric (Appendix A.2.2).

Scaling to black-box LLMs with small surrogates. An obvious question in this setting is whether BAFA performs well even when the audited system is a large black-box LLM like GPT-4.1-mini, while using a much smaller surrogate such as BERT for constraint optimisation. However, passive baselines achieve lower AUEC at very small budgets in Case Study B because the underlying disparity ($\mu_{\Delta\mathrm{ROC\ AUC}} \approx 0.02$–$0.045$) is smaller than in Case Study A, resulting in lower-variance $\Delta\mathrm{AUC}$ estimates at small sample sizes. More importantly, beyond this initial phase, both BAFA variants outperform the baselines in convergence rate (reaching $\epsilon = 0.02$ with $\approx 5 - 41\times$ fewer queries than stratified sampling for case study A ($\approx 41\times$) and B ($\approx 5\times$), and by 250 queries they exhibit reduced variance and achieve AUEC over the first 1,000 queries that is approximately 60% of the baseline AUEC, despite the larger early mean error. In practice, BAFA reduces total AUEC and reaches target precision ϵ with fewer queries in this regime, and replacing the surrogate with DistilBERT increases AUEC by less than 5% (for 3 seeds). We interpret this as consistent with the version space view of (Yan and Zhang, 2022) in that BAFA need not match the black box in parameter space, but must fit queried scores well enough that the constrained optimisation remains feasible and yields a non-trivial interval for $\Delta\mathrm{AUC}$ that shrinks as more informative queries are added. Nevertheless, when the surrogate and audited model are architecturally mismatched, the resulting intervals should be treated as an operational proxy rather than a coverage guarantee.

From failure discovery to quantified uncertainty. We treat auditing as *uncertainty reduction over a target metric*: given a query budget, the goal is not to find the worst failure but to narrow the interval of plausible values for a population-level property such as $\Delta\mathrm{AUC}$, until the remaining uncertainty is small enough to support a defensible claim. This contrasts with red teaming, which takes inspiration from Bayesian sequential black-box testing (Lee et al., 2023) and typically aims to surface as many failures as possible within a fixed budget (Feffer et al., 2025), without estimating prevalence or magnitude. This also clarifies how our approach relates to hypothesis-testing in auditing: Cen and Alur (2024) argue that audits can be framed as hypothesis tests, which is useful for binary compliance decisions under a legal standard. Yet under limited budgets and potential distribution shift, conclusions become sensitive to the chosen threshold and prior assumptions (Juarez et al., 2022). Reporting calibrated uncertainty about the audited quantity is therefore often more informative than a pass/fail certificate, and hypothesis tests can be treated as a downstream decision step, e.g., declaring non-compliance only if the entire uncertainty interval lies above a regulatory standard.

Implications for independent evaluation and continuous monitoring. This framework can support independent evaluators such as NGOs, journalists, and academic auditors in identifying downstream harms under constrained access. When black-box queries are costly, a smaller budget makes it feasible to audit more groups, domains, and languages, and to test targeted hypotheses about where harms may occur (e.g., subgroup-specific false positives that drive unfair moderation). More broadly, the results support continuous monitoring. Instead of running just one benchmark, an auditor can regularly check for disparities, e.g., after model updates, policy changes, or language evolutions, as was called for in hate speech moderation by Tonneau et al. (2025) and Hartmann et al. (2025b). One possible direction for future work is to see auditing as a process of information gain over time, where each new label updates our understanding of fairness and possible distribution shifts. This would help make better use of past audit data and allow for more flexible monitoring.

Understanding contextual implications of input query selection. Active auditing has an additional benefit, namely, that the sequence and composition of queried examples indicate which inputs are selected as most informative under its constraints, offering a potential form of interpretability (as in (Phillips et al., 2018)). BAFA-BO builds on this by using a LoRA surrogate together with a query-diversity signal (as in (Lee et al., 2023)). This combination can make it even more interpretable for understanding selection patterns. Future work should build on this to characterise which regions of the input space different selection rules emphasise, for instance, borderline cases, specific linguistic patterns (e.g., AAE or counter-speech (Sap et al., 2019)), identity tokens or particular subpopulations. Such analyses could guide further qualitative investigation and stakeholder review of the sampled content.

Generalisability beyond ROC AUC difference. Lastly, while we instantiate our framework for ΔAUC, the broader idea is that active, query-efficient auditing can be applied whenever a black-box system exposes a reliable signal that can be turned into a scorable objective (differentiable or well-approximated by a smooth surrogate), enabling optimisation and uncertainty-aware selection. This covers other group metrics (e.g., TPR/FPR gaps at fixed thresholds, equalised odds, see Gallegos et al. (2024)) and extends to performance (Ribeiro et al., 2020), privacy audits (Staufer, 2025) and robustness and safety audits (Rauba et al., 2025), as well as benchmarking (Liang et al., 2023).

7 Conclusion

We presented BAFA, a query-efficient framework for auditing group fairness of black-box language models under realistic access and budget constraints. Across two auditing scenarios – hate speech detection and profession inference – BAFA consistently reduced the number of required queries by one order of magnitude compared to sampling baselines and ablations, while achieving lower estimation error and improved stability at moderate budgets. Conceptually, our results support viewing auditing as uncertainty estimation over a target metric rather than failure discovery or one-shot benchmarking. While BAFA does not resolve downstream harms or replace qualitative evaluation, it provides a practical measurement tool for making independent fairness audits with limited access more feasible, interpretable, and precise for black-box LLMs.

Limitations

Surrogate model choice and the computational-precision trade-off We chose BERT-base as our surrogate model to keep computational costs low, which is important for independent auditors like civil society groups, journalists, and academic researchers who often have limited resources. There is a trade-off, though: the method works best when the surrogate model is similar to the black-box system being audited (though our ablations found only marginal differences when switching to DistillBERT). If auditors know the system's architecture and have more resources, they can use a larger or better-matched surrogate, such as GPT-2 for auditing GPT-3, or RoBERTa-large for more complex tasks. Future work should especially try out GPT-2 or GPT-3 for the GPT-4.1-mini audit as architectures are the same and, thus, could lead to more accurate results and faster convergence. However, such experiments are out of scope for this work due to the focus on independent audits. In our experiments for Case Study B, we show that BAFA still performs very well even when the surrogate and target architectures do not match exactly.

Computational and resource intensity. A key limitation is that our end-to-end pipeline is resource intensive as it requires repeated optimisation steps within the loop. This is costly in wall-clock time and GPU usage, especially when scaling to many seeds, many groups, or frequent monitoring (see Appendix section A.4.5 for a detailed analysis of computational resources needed). This directly conflicts with our motivating goal of enabling resource-efficient auditing for independent evaluators. However, we think that substantial speedups are likely feasible. Promising directions include engineering improvements (e.g., caching/more efficient data pipelines), algorithmic warm-starting across rounds, more efficient batching strategies, and hybrid protocols that switch to simpler sampling once the interval is already narrow but a careful study of these system-level trade-offs is unfortunately out of scope for this work.

From a research prototype to an auditor-facing tool. While BAFA demonstrates the feasibility of query-efficient, uncertainty-aware auditing in con-

trolled experimental settings, it is not yet a finished tool that can be readily deployed by independent auditors in practice. Turning BAFA into a practical auditing tool would therefore require integrating the needs and requirements of stakeholders and users, including support for multiple evaluation metrics (fairness-related or otherwise), transparent uncertainty reporting, and simple mechanisms for updating datasets and managing query budgets. A promising direction is the development of human-centered interfaces that allow auditors to configure audits through intuitive interactions (e.g., selecting metrics, uploading or modifying datasets, and issuing queries via clicks or drag-and-drop with uncertainty visualization). We see BAFA as a methodological building block toward such systems, but significant design, engineering, and participatory work remains to translate it into a robust and usable auditing infrastructure.

From metric gaps to downstream harms and the limits of "certificates". Finally, fairness metrics (including bounded disparity estimates) are only proxies for real-world harm. Connecting a measured gap to downstream impacts requires context interpretations: whom the system affects, how it is used, and what policies and incentives shape outcomes (Blodgett et al., 2020). In many cases, quantitative disparity estimates alone will not surface the most important harms (Raji et al., 2021). We therefore see metric-based auditing as most useful when paired with complementary methods such as qualitative methods, stakeholder engagement, and case-based human-centered evaluations, including affected users' experiences (Liu et al., 2025).

Our uncertainty bounds can also be read as a kind of certificate but only for the audited metric under the audit distribution and assumptions, and only at a particular snapshot in time. They should not be mistaken for a guarantee that the overall system is safe, fair, or non-harmful. Although the model might have tight bounds and satisfy the fairness criteria metric, the model can still cause substantial harm that is not captured by the chosen metric. This is another reason for us to claim that thinking of auditing from an uncertainty perspective rather than a hypothesis-testing and compliance perspective could be a step towards less reliance on technical fairness metrics.

Ethical Considerations

Responsible use and the risk of "ethics washing". Our work is meant to make fairness auditing more accessible to under-resourced groups, such as civil society organisations, journalists, academic researchers and generally for independent auditing organisations. Still, like all auditing tools, the tool can be misused to give a false sense of accountability without real systemic change (Raji et al., 2020; Hartmann et al., 2025a) or in the case of red teaming "security theatre" (Feffer et al., 2025). Companies that have a self-interest in demonstrating surface compliance might only audit metrics where they perform well, or use our method to give false reassurance. This is why we want to stress that BAFA is a measurement tool, not a solution to algorithmic harm. Query-efficient auditing helps detect disparities, but fixing them needs organisational commitment, policy changes, and involvement from affected communities in making decisions about remedies.

LLM-based Tools. We used LLM-based assistance tools in a limited way during manuscript preparation and implementation. GitHub Copilot was used for code completion and minor refactoring, and Claude was used to suggest alternative phrasings and polish LaTeX formatting (for example, the table layout) in appendix sections. All algorithmic design decisions, experimental implementation and execution, data analysis, and substantive writing were carried out by the authors, and we verified any AI-assisted edits for correctness.

Acknowledgments

This work was partially supported by the German Federal Ministry of Research, Technology and Space (BMFTR) — Nr. 16DII144 and the Deutsche Forschungsgemeinschaft (DFG), project number 460135501, NFDI 29/1 "MaRDI – Mathematische Forschungsdateninitiative". This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

References

Shivani Agarwal. 2013. Surrogate regret bounds for the area under the roc curve via strongly proper losses. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Ma-*

chine Learning Research, pages 338–353, Princeton, NJ, USA. PMLR.

Shivani Agarwal, Thore Graepel, Ralf Herbrich, Sariel Har-Peled, Dan Roth, and Michael I Jordan. 2005. Generalization bounds for the area under the roc curve. *Journal of Machine Learning Research*, 6(4).

Ayoub Ajarra, Bishwamittra Ghosh, and Debabrota Basu. 2024. Active fourier auditor for estimating distributional properties of ml models. *arXiv preprint arXiv:2410.08111*.

Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. Ai auditing: The broken bus on the road to ai accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 612–643. IEEE.

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Daniel Borkan, Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019a. Limitations of pinned auc for measuring unintended bias. *Preprint*, arXiv:1903.02088.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019b. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25. Association for Computational Linguistics.

Sarah H. Cen and Rohan Alur. 2024. From Transparency to Accountability and Back: A Discussion of Access and Evidence in AI Auditing. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–14. ACM.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Pieter Delobelle, Giuseppe Attanasio, Debora Nozza, Su Lin Blodgett, and Zeerak Talat. 2024. Metrics for what, metrics for whom: Assessing actionability of bias evaluation metrics in nlp. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 21669–21691, Singapore. Association for Computational Linguistics. Joint first authors: Delobelle and Attanasio.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, pages 4171–4186.

Michael Feffer, Anusha Sinha, Wesley H. Deng, Zachary C. Lipton, and Hoda Heidari. 2025. *Red-Teaming for Generative AI: Silver Bullet or Security Theater?*, page 421–437. AAAI Press.

Jose Gallego-Posada, Juan Ramirez, Meraj Hashemizadeh, and Simon Lacoste-Julien. 2025. Cooper: A Library for Constrained Optimization in Deep Learning. *arXiv preprint arXiv:2504.01212*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics. Allen Institute for AI and University of Washington.

Kate Glazko, Yusuf Mohammed, Ben Kosa, Venkatesh Potluri, and Jennifer Mankoff. 2024. Identifying and Improving Disability Bias in GPT-Based Resume Screening. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 687–700. ACM.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Mu noz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics. Equal contribution: Goldfarb-Tarrant, Marchant, Muñoz Sánchez, Pandya.

David Hartmann, José Renato Laranjeira De Pereira, Chiara Streitbörger, and Bettina Berendt. 2025a. Addressing the regulatory gap: Moving towards an EU AI audit ecosystem beyond the AI Act by including civil society. *AI and Ethics*.

David Hartmann, Amin Oueslati, Dimitri Staufer, Lena Pohlmann, Simon Munzert, and Hendrik Heuer. 2025b. Lost in moderation: How commercial content moderation apis over- and under-moderate group-targeted hate speech and linguistic variations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2025. Evaluating Bias in LLMs for Job-Resume Matching: Gender, Race, and Education. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 672–683. Association for Computational Linguistics.

Marc Juarez, Samuel Yeom, and Matt Fredrikson. 2022. Black-box audits for group distribution shifts. *arXiv preprint arXiv:2209.03620*.

Deokjae Lee, JunYeong Lee, Jung-Woo Ha, Jin-Hwa Kim, Sang-Woo Lee, Hwaran Lee, and Hyun Oh Song. 2023. Query-efficient black-box red teaming via bayesian optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 32 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Published 23 Aug 2023. Also available as arXiv:2211.09110.

Yu Lu Liu, Wesley Hanwen Deng, Michelle S. Lam, Motahhare Eslami, Juho Kim, Q. Vera Liao, Wei Xu, Jekaterina Novikova, and Ziang Xiao. 2025. Human-centered evaluation and auditing of language models. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Pranav Maneriker, Codi Burley, and Srinivasan Parthasarathy. 2023. Online fairness auditing through iterative refinement. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 1665–1676, New York, NY, USA. Association for Computing Machinery.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2024. Auditing large language models: a three-layered approach. *AI and Ethics*, 4:1085–1115.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics. Equal contribution.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Richard Phillips, Kyu Hyun Chang, and Sorelle A. Friedler. 2018. Interpretable active learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 49–61. PMLR.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. Ai and the everything in the whole wide world benchmark. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*. Track on Datasets and Benchmarks.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 33–44, New York, NY, USA. Association for Computing Machinery.

Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel E. Ho. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. *arXiv preprint*. ArXiv:2206.04737 [cs].

Bashir Rastegarpanah, Krishna P Gummadi, and Mark Crovella. 2021. Auditing black-box prediction models for data minimization compliance. In *Advances in Neural Information Processing Systems*, volume 34, pages 10001–10014. NeurIPS.

Paulius Rauba, Qiyao Wei, and Mihaela van der Schaar. 2025. Statistical hypothesis testing for auditing robustness in language models. *arXiv preprint arXiv:2506.07947*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual hatecheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Harvineet Singh, Fan Xia, Mi-Ok Kim, Romain Pirracchio, Rumi Chunara, and Jean Feng. 2023. A brief tutorial on sample size calculations for fairness audits. *Preprint*, arXiv:2312.04745.

Dimitri Staufer. 2025. What should LLMs forget? quantifying personal data in LLMs for right-to-be-forgotten requests. In *Proceedings of the 7th Workshop on eXplainable Knowledge Discovery in Data Mining (XKDD)*. Co-located with ECML PKDD 2025, Porto, Portugal.

Manuel Tonneau, Diyi Liu, Niyati Malhotra, Scott A. Hale, Samuel P. Fraiberger, Victor Orozco-Olvera, and Paul Röttger. 2025. Hateday: Insights from a global hate speech dataset representative of a day on twitter. In *Proceedings of the 2025 Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ze Wang, Zekun Wu, Xin Guan, Michael Thaler, Adriano Koshiyama, Skylar Lu, Sachin Beepath, Ediz Ertekin, and Maria Perez-Ortiz. 2024. JobFair: A Framework for Benchmarking Gender Hiring Bias in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3227–3246. Association for Computational Linguistics.

Tom Yan and Chicheng Zhang. 2022. Active fairness auditing. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24929–24962. PMLR.

Juliette Zaccour, Reuben Binns, and Luc Rocher. 2025. Access denied: Meaningful data access for quantitative algorithm audits. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Formal Problem Setup and Version Space

Black-box Model and Data. We assume a black-box model $h^\star : \mathcal{X} \to \mathbb{R}$ that returns scores for inputs $x \in \mathcal{X}$, where $\mathcal{X}$ denotes the input space (e.g., text documents). Given labeled data $\mathcal{D} = \{(x_i, y_i, g_i)\}_{i=1}^N$ with binary label $y_i \in \{0, 1\}$ and protected group attribute $g_i \in \{0, 1\}$, the goal is to estimate a fairness measure μ. In this work, we focus on the group fairness disparity measured by the *Area under the ROC curve (AUC) difference*:

$$\Delta_{\text{AUC}}(h^\star) = \text{AUC}_{g=0}(h^\star) - \text{AUC}_{g=1}(h^\star),$$

where $\text{AUC}_g(h) = \mathbb{P}(h(X_g^+) > h(X_g^-))$ with $(X^+, X^-) \sim \mathcal{D}_{X|Y=1,G=g} \times \mathcal{D}_{X|Y=0,G=g}$ representing independent draws from the positive and negative class distributions within group g.

Audit Objective. Given a query budget T, we seek an estimator $\widehat{\Delta}_{\text{AUC}}$ that is ϵ-accurate with high probability:

$$\mathbb{P}\left(\left|\widehat{\Delta}_{\text{AUC}} - \Delta_{\text{AUC}}(h^\star)\right| \le \epsilon\right) \ge 1 - \delta,$$

while minimizing the number of queries $q \le T$. We assume access to ground-truth labels and group attributes for the audit pool, but only black-box query access to $h^\star$—we cannot inspect model internals, parameters, or training data.

Queried Set and Surrogate Hypothesis Class. At audit round t, let $S_t \subseteq \mathcal{D}$ denote the set of examples queried so far, where each $(x_i, y_i, g_i) \in S_t$ is augmented with its black-box score $s_i^\star = h^\star(x_i)$. We maintain a surrogate hypothesis class $\mathcal{H}$ (in our case, a parameterized neural network family such as BERT-based classifiers) and define the *version space* as the set of surrogate hypotheses consistent with the observed queries:

Version Space. Given a tolerance parameter $\lambda > 0$, the λ-approximate version space is:

$$\mathcal{H}_\lambda(S_t) = \left\{ h \in \mathcal{H} : \begin{matrix} |h(x_i) - s_i^\star| \le \lambda, \\ \forall (x_i, s_i^\star) \in S_t \end{matrix} \right\}.$$

This set contains all surrogate models that approximate the black-box scores on queried examples within tolerance λ. As more examples are queried, the version space $\mathcal{H}_\lambda(S_t)$ becomes increasingly constrained, and the range of possible fairness values $\mu(h)$ for $h \in \mathcal{H}_\lambda(S_t)$ narrows.

A.2 Implementation Details (BAFA)

This section specifies the mechanics of BAFA: (i) how we compute the certificate interval via constrained optimisation, and (ii) how we implement active query selection, including distribution regularisation, diversity, and BO. Throughout, the audited system is treated as a black box; BAFA only observes scalar scores returned by a query API. The pseudo-code is presented in Algorithm 2.

A.2.1 Audit pool, interfaces, and invariants

Audit pool. BAFA operates on a fixed audit pool $\mathcal{U} = \{(x_i, g_i, y_i, \text{id}_i)\}_{i=1}^N$, where x_i is the input (text), g_i is the protected attribute, y_i is the ground-truth label used to define the fairness metric, and id_i is a deterministic identifier. We treat $\mathcal{U}$ as immutable and never reindex after construction.

Black-box interface. The audited system is accessed only via a scoring interface

$$h^\star(x) \to s^\star \in [0, 1],$$

returning a scalar score for the positive class (toxicity / one-vs-rest occupation probability). We maintain an incrementally growing queried set $S_t \subset \mathcal{U}$, where each queried point is augmented with its black-box score $s_i^\star = h^\star(x_i)$. All selection and logging is keyed by id to prevent accidental re-querying and to keep cached artifacts (scores, embeddings) aligned to $\mathcal{U}$.

Fairness estimator on a queried set. Given a queried set S_t with scores $\{s_i^\star\}$, we compute the empirical group AUCs and their difference

$$\widehat{\Delta}_{\text{AUC}}(S_t) = \widehat{\text{AUC}}_{g=0}(S_t) - \widehat{\text{AUC}}_{g=1}(S_t),$$

using the standard ROC-AUC estimator within each group. If a group in S_t contains only one label class, the group AUC is undefined; we then treat $\widehat{\Delta\text{AUC}}(S_t)$ as missing for that time step (this affects only very small budgets in heavily imbalanced strata).

A.2.2 Bound step: constrained ERM with Cooper

At each round t, BAFA computes an uncertainty interval $[\mu_{\min}^t, \mu_{\max}^t]$ for the target metric $\mu(\cdot)$ by solving two constrained optimisation problems over a surrogate hypothesis class $\mathcal{H}$.

Version space constraint. Let S_t be the queried set and λ be the score-tolerance parameter. We define an approximate version space

$$\mathcal{H}_\lambda(S_t) = \{h \in \mathcal{H} : |h(x_i) - s_i^\star| \le \lambda, \forall (x_i, \cdot) \in S_t\}.$$

In practice, we enforce these constraints via a differentiable Lagrangian formulation using `cooper` (Gallego-Posada et al., 2025), which maintains primal parameters (surrogate weights) and dual variables (Lagrange multipliers) and performs constrained updates.

Extremal hypotheses and certificate. We compute two feasible hypotheses by extremising the fairness objective:

$$h^t_{\max} \in \arg\max_{h \in \mathcal{H}_\lambda(S_t)} \mu(h),$$

$$h^t_{\min} \in \arg\min_{h \in \mathcal{H}_\lambda(S_t)} \mu(h).$$

The resulting certificate interval is

$$\mu^t_{\max} := \mu(h^t_{\max}), \qquad \mu^t_{\min} := \mu(h^t_{\min}).$$

We report the midpoint estimate $\hat{\mu}_t := (\mu^t_{\min} + \mu^t_{\max})/2$ and interpret the half-width $(\mu^t_{\max} - \mu^t_{\min})/2$ as the current uncertainty radius.

Objective implementation. To enable gradient-based optimisation, we implement $\mu(h)$ using a smooth proxy of ΔAUC that is consistent with the empirical AUC difference. Concretely, we express each group AUC as a U-statistic over positive–negative pairs and replace the indicator $\mathbb{1}_{[h(x^+)>h(x^-)]}$ with a sigmoid comparator $\sigma((h(x^+) - h(x^-))/\tau)$ (temperature $\tau > 0$). This yields a differentiable approximation to ΔAUC used in the inner optimisation; evaluation and reporting still use the standard ROC-AUC estimator on black-box scores.

Calibration of uncertainty intervals We assess empirical calibration of BAFA's uncertainty interval $[\mu^t_{\min}, \mu^t_{\max}]$ by measuring (i) **coverage**, i.e., whether the ground-truth disparity Δ_{true} lies within $[\mu^t_{\min}, \mu^t_{\max}]$, and (ii) **bound violation**, defined as $\max\{0, \mu^t_{\min} - \Delta_{\text{true}}, \Delta_{\text{true}} - \mu^t_{\max}\}$ (in ΔAUC ROC points). Figure 6 visualizes the violation distributions and Tables 2–3 summarize results. On Jigsaw, intervals are well calibrated with near-zero violations, consistent with stronger surrogate–black-box alignment; on Bias-in-Bios, coverage is lower, but violations are typically small (mean $< \varepsilon$ for strict $\varepsilon = 0.02$), so interval width remains a useful operational *proxy* for uncertainty even when it should not be interpreted as a formal coverage guarantee.

Signals exposed to the selector. The selector uses the two extremal hypotheses to score candidates on $U \setminus S_t$:

$$h^t_{\min}(x), \qquad h^t_{\max}(x).$$

These scores are used to compute disagreement and (optionally) expected-width reduction signals for active sampling.

A.2.3 Selection step: ordered sampling rules

BAFA selects the next batch of queries using `AuditSelector` (`selection.py`). Let U denote the audit pool and S_t the currently queried set. At each round we form the unqueried candidate set $U \setminus S_t$. For efficiency, the runner may additionally subsample a candidate pool of size M from $U \setminus S_t$ before scoring (this changes runtime but not the definition of any strategy).

Random and stratified baselines. `random` samples k points uniformly without replacement from $U \setminus S_t$. `stratified` performs proportional stratified sampling over strata. In our experiments, the seed set S_0 is stratified over (g, y) when labels are available; subsequent stratified batches are stratified over g (and optionally (g, y) when required by the evaluation protocol). Formally, for a requested sample size n, the stratified sampler allocates

$$n_s \approx \left\lceil n \cdot \frac{|U_s|}{|U|} \right\rceil \quad \text{for each stratum } s,$$

samples n_s points uniformly without replacement from each stratum subset, and concatenates them.

BAFA-Disagreement. Disagreement is defined directly from the certificate endpoints, as in Algorithm 2:

$$\text{dist}(x) = \left| h^t_{\max}(x) - h^t_{\min}(x) \right|.$$

The selector assigns each candidate a final score $s_t(x)$ (defined below) and queries the top-k candidates.

BAFA-BO (disagreement-anchored BO). `bo` implements a stabilised variant of Bayesian optimisation (BO) in feature space. The key design choice is that BO is bounded and anchored: it does

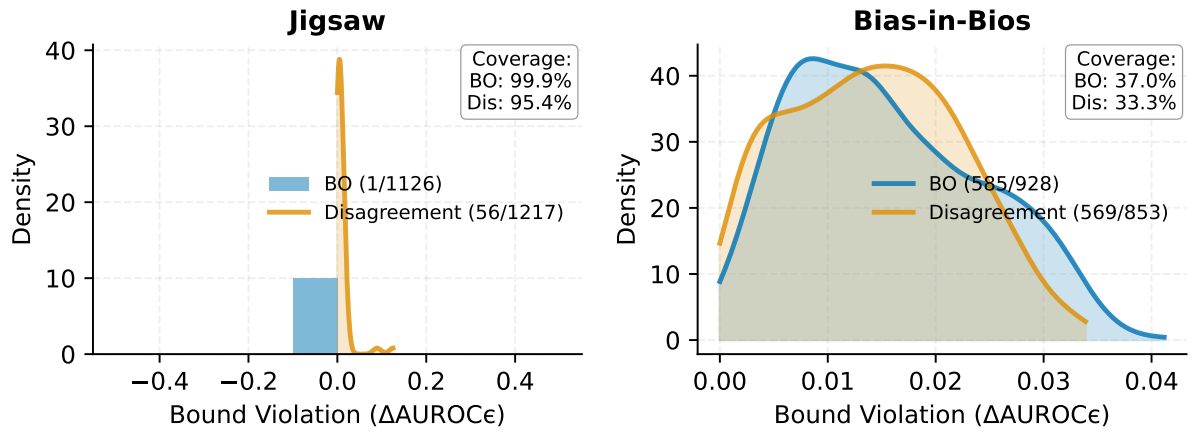

Figure 6: **Bound violation distributions.** A bound violation is the amount by which Δ_{true} falls outside BAFA's uncertainty interval $[\mu_{\min}, \mu_{\max}]$ (zero if inside). Jigsaw shows near-zero violations and high empirical coverage, whereas Bias-in-Bios exhibits more frequent violations but typically small magnitude.

Table 2: Uncertainty Calibration: Coverage and Bound Violations

Dataset	Strategy	Coverage	Bound Violation	Median
Jigsaw	BO	99.9%	0.0000 [0.0000, 0.0000]	0.0000
	Disagreement	95.4%	0.0004 [0.0002, 0.0007]	0.0000
Bias-in-Bios	BO	37.0%	0.0098 [0.0091, 0.0104]	0.0073
	Disagreement	33.3%	0.0095 [0.0088, 0.0101]	0.0076

Coverage: Percentage of iterations where $\Delta_{\text{true}} \in [\mu_{\min}, \mu_{\max}]$. **Bound Violation**: Mean distance (in ΔAUROC points) by which Δ_{true} falls outside $[\mu_{\min}, \mu_{\max}]$, with 95% confidence intervals computed via bootstrap (10,000 resamples). Higher coverage and lower violations indicate better empirical calibration.

not replace the certificate-derived informativeness signal, but provides a secondary exploration term whose influence is ramped in gradually.

We construct a feature vector $\phi_t(x)$ by concatenating: (i) $\text{dist}(x)$, (ii) an optional gradient feature $\text{grad}(x)$ (if provided by `gradient_fn`), and (iii) an optional surrogate embedding $\phi_{\text{sur}}(x)$ (e.g., BERT [CLS]) produced by `surrogate_feat_fn`. All features are sanitised (NaN/Inf $\to$ 0).

We then fit a Gaussian Process in feature space and compute a UCB acquisition score:

$$\text{acq}_t(x) = \mu_{\text{GP}}(\phi_t(x)) + \beta\, \sigma_{\text{GP}}(\phi_t(x)).$$

To avoid numerical dominance, we z-score acq_t across the candidate pool and squash it into $[0, 1]$ via a clipped logistic transform, yielding $\text{acq}_{t,01}(x)$. The mixed informativeness score is

$$\text{comb}_t(x) = (1 - \lambda_t)\, \text{dist}(x) + \lambda_t\, \text{acq}_{t,01}(x),$$

where λ_t follows a warm-up-and-ramp schedule ($\lambda_t = 0$ early and $\lambda_t \le \lambda_{\max}$ later). This implements the "anchor vs. stabiliser" design: disagreement remains the primary driver while BO contributes a bounded exploration term.

BO state management. The runner maintains a BO dataset $(X_{\text{BO}}, y_{\text{BO}})$ over time (feature vectors and observed utility). In BAFA, the utility y_t is a per-query proxy for audit progress: realised certificate width reduction attributable to previously queried points, i.e., $y_t = (W^{t-1} - W^t)/|Q_{t-1}|$ as used in Algorithm 2. To prevent stale behaviour, we refit the GP whenever the BO dataset size changes; the selector caches the fitted GP and tracks the training-set size for refit decisions.

A.2.4 Regularisation in the selector

Regularisation acts only in the selection module; the certificate computation is unchanged. We use three complementary mechanisms.

(1) Distribution matching weights. Active strategies may induce selection bias by oversampling particular group–label strata. To control drift between the queried distribution $p_{S_t}(g, y)$ and the pool distribution $p_U(g, y)$, we compute per-stratum

Table 3: Detailed Uncertainty Diagnostics

Dataset	Strategy	n	Coverage	Pearson r	Spearman ρ
Jigsaw	BO	1226	99.9%	0.853	0.530
	Disagreement	1217	95.4%	0.738	0.324
Bias-in-Bios	BO	1228	37.0%	0.395	0.203
	Disagreement	1253	33.3%	0.442	0.255

Pearson and Spearman correlations measure the relationship between predicted interval width ($\mu_{\max} - \mu_{\min}$) and realized absolute error $|\hat{\mu} - \Delta_{\text{true}}|$. Strong positive correlations (Jigsaw) indicate that wider intervals reliably predict larger errors, while weak correlations (Bias-in-Bios) indicate poorer calibration.

weights and multiply them into the selection score. Each candidate (x, g, y) receives a weight

$$w_{(g,y)} = 1+\alpha_t \left(\min\left(c_{\max}, \frac{p_U(g,y)}{\max(p_{S_t}(g,y), \varepsilon)} \right) - 1 \right),$$

matching the form given in Algorithm 2, with cap $c_{\max}$ to avoid extreme weights in rare strata. α_t follows a warm-up-and-ramp schedule. If a stratum is absent, we default to $w_{(g,y)} = 1$.

(2) Diversity regularisation (MMR-style batch construction). For BO-based strategies we apply an MMR-style penalty during greedy top-k selection to avoid near-duplicates. Given current selected set Q_t, we score a remaining candidate x by

$$s_t^{\text{div}}(x) = s_t(x) - \gamma \max_{x' \in Q_t} \text{sim}\big(\phi_t(x), \phi_t(x')\big),$$

where sim is cosine similarity of ℓ_2-normalised features. This improves coverage of the candidate space at fixed batch size.

(3) Optional BO restriction to high-disagreement regions. Optionally, BO mixing is applied only within a high-disagreement subset defined by a quantile threshold on $\text{dist}(x)$. Outside this region, the selector defaults to the anchor signal. This is a conservative safeguard when the GP signal is unreliable.

Final score. For disagreement / EWR strategies, the score is $s_t(x) = \text{dist}(x) \cdot w_{(g,y)}$. For BO strategies, the score is $s_t(x) = \text{comb}_t(x) \cdot w_{(g,y)}$, followed by diversity-aware batch selection.

A.2.5 Diagnostics, numerical stability, and reproducibility

Diagnostics. The selector records per-round buffers over the candidate pool (raw informativeness, acquisition values, final scores, selected features, and selected IDs). These logs support post-hoc analyses of what the auditor considered informative (e.g., boundary cases vs. under-covered strata) and enable clean ablations that remove individual regularisers while keeping the rest fixed.

Numerical stability. We apply defensive guards throughout selection and BO: clipping exponentials in logistic transforms, adding ε to standard deviations in z-scoring, sanitising NaN/Inf values in features and scores, and capping ratio-based distribution weights. These guards matter at small budgets where $p_{S_t}(g, y)$ can be near zero and where GP fits can be ill-conditioned.

Index/ID invariants. A critical implementation invariant is that all sampling and concatenation preserves the original id keys from U. We never reset indices after pool construction, and we compute "already queried" sets only via IDs. This prevents subtle failures where embeddings, cached scores, or selection masks drift out of alignment with U.

A.3 Baselines and Ablations: Sampling Rules and Estimators

All methods operate on the same audit pool $\mathcal{U}$ and differ only in the ordered sampling rule that selects the next batch of black-box queries. Let S_t denote the queried set after t total queries (including the seed set). Each method outputs a trajectory of fairness estimates $\widehat{\Delta}_{\text{AUC}}(S_t)$ using the same estimator defined in Appendix A.2.1.

Common initialisation. All methods start from the same seed set S_0, obtained by stratified sampling with size k_{init} (over (g, y) when labels are available), followed by querying the black-box to attach scores.

A.3.1 Passive sampling baselines

Random sampling. At each round, sample k points uniformly without replacement from $\mathcal{U} \setminus S_t$.

Stratified sampling. Stratified sampling preserves representativeness of protected groups (and

Algorithm 2 BAFA: Bounded Active Fairness Auditing with C-ERM (expanded)

Require: Audit pool U, black-box $h^\star$, budget T, tolerance λ, batch size k, threshold ϵ, selector $\Pi \in \{\text{dis}, \text{bo}\}$

$S_0 \leftarrow$ stratified seed set; query $h^\star$ to attach scores $\{s_i^\star\}$; init. BO dataset $(X_{\text{BO}}, y_{\text{BO}}) \leftarrow (\emptyset, \emptyset)$
for $t = 0, 1, 2, \ldots, T$ **do**
  // S1: Certificate
  Solve $h^t_{\max} \leftarrow \arg\max_{h \in \mathcal{H}_\lambda(S_t)} \mu(h)$
  Solve $h^t_{\min} \leftarrow \arg\min_{h \in \mathcal{H}_\lambda(S_t)} \mu(h)$
  $[\mu^t_{\min}, \mu^t_{\max}] \leftarrow [\mu(h^t_{\min}), \mu(h^t_{\max})]$
  if $(\mu^t_{\max} - \mu^t_{\min})/2 \le \epsilon$ **then**
    return $\hat{\mu}_t = (\mu^t_{\min} + \mu^t_{\max})/2$
  end if
  // S2: Active query selection
  for each $x \in U \setminus S_t$ **do**
    $\text{dist}_t(x) \leftarrow |h^t_{\max}(x) - h^t_{\min}(x)|$
  end for
  $r_{(g,y)} \leftarrow \min(c_{\max}, p_U(g,y)/p_{S_t}(g,y))$
  $w_{(g,y)} \leftarrow 1 + \alpha_t(r_{(g,y)} - 1)$
  if $\Pi = \text{dis}$ **then**
    $s_t(x) \leftarrow \text{dist}(x) \cdot w_{(g,y)}$; $Q_t \leftarrow$ top-k by s_t
  else if $\Pi = \text{bo}$ **then**
    $\phi_t(x) \leftarrow [\text{dist}(x), \text{grad}(x), \phi_{\text{sur}}(x)]$
    $\text{acq}_t(x) \leftarrow \mu_{\text{GP}}(\phi_t) + \beta\sigma_{\text{GP}}(\phi_t)$
    $\text{acq}_{t,01} \leftarrow \sigma(\text{z-score}(\text{acq}_t))$
    $\text{comb}_t(x) \leftarrow (1 - \lambda_t)\text{dist}(x) + \lambda_t \text{acq}_{t,01}(x)$
    $s_t(x) \leftarrow \text{comb}_t(x) \cdot w_{(g,y)}$
    $Q_t \leftarrow$ MMR top-k with penalty $\gamma \max_{x' \in Q_t} \text{sim}(\phi_t(x), \phi_t(x'))$
  end if
  Query $h^\star$ on Q_t; update $S_{t+1} \leftarrow S_t \cup Q_t$
  if $\Pi = \text{bo}$ **and** $t > 0$ **then**
    Append $(\phi_{t-1}(x), y_t)$ for $x \in Q_{t-1}$ to $(X_{\text{BO}}, y_{\text{BO}})$, with $y_t = (W^{t-1} - W^t)/|Q_{t-1}|$
  end if
end for

optionally group–label strata). For a requested sample size n, we allocate approximately proportional quotas n_s per stratum s and sample uniformly within each stratum without replacement. This is a strong passive baseline in our setting because it controls group-marginal drift while remaining label-agnostic beyond the strata definition.

Baseline Bounds In the following, we derive a bound on the number of samples required so that

$$\mathbb{P}\left(\left|\widehat{\Delta}_{\text{AUC}} - \Delta_{\text{AUC}}\right| \ge \epsilon\right) \le \delta$$

for given tolerance thresholds ϵ and total failure probability δ using McDiarmid's inequality for AUC (Agarwal et al., 2005). To start with, application of the union bound to the left-hand side yields

$$\mathbb{P}\left(\left|\widehat{\Delta}_{\text{AUC}} - \Delta_{\text{AUC}}\right| \ge \epsilon\right) \le \sum_{g \in \{0,1\}} \mathbb{P}\left(\left|\widehat{\text{AUC}}_g - \text{AUC}_g\right| \ge \epsilon/2\right).$$

Now we may apply McDiarmid's inequality to each summand to obtain

$$\mathbb{P}\left(\left|\widehat{\Delta}_{\text{AUC}} - \Delta_{\text{AUC}}\right| \ge \epsilon\right) \le \sum_{g \in \{0,1\}} 2\exp\left(-\frac{2m_g n_g (\epsilon/2)^2}{m_g + n_g}\right),$$

where m_g and n_g describe the number of positive and negative samples, for each group g, respectively. Assuming a balanced sampling of positive and negative as well as between-group samples $m_g \approx n_g \approx n$ we may solve for n, resulting in the following lower bound on the number of samples:

$$\frac{8}{\epsilon^2} \log\left(\frac{4}{\delta}\right) \le n.$$

Power sampling. Power sampling prioritises boundary-adjacent points using the score-uncertainty proxy $u(x) = p(x)(1 - p(x))$ with $p(x) = h^\star(x)$. It samples points proportionally to $u(x)^\gamma$:

$$\Pr(x_i \text{ selected}) \propto \big(p_i(1 - p_i)\big)^\gamma.$$

This can accelerate estimation of ranking-based metrics but can also concentrate queries in narrow regions of the input space and induce selection bias.

A.3.2 BO baseline (sampling-only)

Bayesian optimisation baseline. The BO baseline is a sampling rule that fits a GP on text embeddings and selects points with a standard BO acquisition function (e.g., EI/UCB). Crucially, this baseline does not use BAFA's certificate endpoints

and does not optimise interval shrinkage. We include it as a representative embedding-based BO heuristic to contrast with BAFA-BO, where BO is anchored to certificate-derived informativeness and used only as a bounded stabiliser.

A.3.3 C-ERM ablation (certificate without active selection)

C-ERM-only ablation (passive acquisition, certificate estimator). To isolate the effect of active selection from the effect of certificate-based estimation, we consider a C-ERM ablation that removes active selection entirely: (i) acquire samples using a passive rule (stratified per round), (ii) after each acquisition, run C-ERM twice to compute $[\mu^t_{\min}, \mu^t_{\max}]$, (iii) report the midpoint $\hat{\mu}_t$ and width. This ablation keeps BAFA's estimator but removes certificate-informed query allocation.

A.4 Experimental Details

A.4.1 Case Study A: CivilComments Black-Box Scoring & Reproducibility

This case study audits racial disparities in hate speech detection on the CivilComments dataset (Borkan et al., 2019b). We treat a fine-tuned Transformer classifier as a black-box scorer $h^\star$ and estimate the fairness target $\Delta\mathrm{AUC}$ between dominant and marginalized identity groups under limited query budgets.

Dataset. We use the CivilComments dataset from the Jigsaw Unintended Bias in Toxicity Classification benchmark. The dataset contains user-generated comments from English-language news sites annotated for toxicity and multiple identity targets. We focus on a binary group comparison between the dominant group (`white`) and the marginalized group (`black`). After filtering for valid group labels and ground-truth toxicity annotations, the audit pool $\mathcal{U}$ contains approximately 50,000 comments. Each example is assigned a deterministic identifier based on its index in the filtered dataset.

Black-box model. The black-box $h^\star$ is a HateBERT model (`GroNLP/hateBERT`) fine-tuned on the SBIC dataset (Sap et al., 2020). The model is trained with a single-logit classification head and outputs a real-valued toxicity score. During fine-tuning, we inject systematic bias by stochastically flipping toxicity labels with fixed, group-conditional probabilities. Labels associated with the marginalized group (`black`) are flipped with substantially higher probability than those associated with the dominant group (`white`), while all randomness is controlled via fixed seeds. This procedure induces a stable ground-truth disparity of approximately $\Delta\mathrm{AUC} \approx 0.14$, with higher AUC for the white group.

Black-box inference. At audit time, the model is treated as a black box and queried only via its scoring interface. For each input comment x_i, the black-box returns a toxicity score $s^\star_i \in [0, 1]$, obtained by applying a sigmoid to the model's output logit. Inference is deterministic, with the model fixed in evaluation mode and no stochastic decoding.

Fairness metric. We compute ROC AUC separately for the dominant and marginalized groups:

$$\mathrm{AUC}_{\text{white}} = \mathrm{AUC}(\{s^\star_i, y_i\}_{\text{group=white}})$$

$$\mathrm{AUC}_{\text{black}} = \mathrm{AUC}(\{s^\star_i, y_i\}_{\text{group=black}}),$$

where $y_i \in \{0, 1\}$ denotes the ground-truth toxicity label. The target fairness metric is the difference

$$\Delta\mathrm{AUC} = \mathrm{AUC}_{\text{white}} - \mathrm{AUC}_{\text{black}}.$$

This $\Delta\mathrm{AUC}$ is the quantity estimated by the active auditing pipeline in the main paper.

Caching and black-box interface. Unlike Case Study B, scores are not cached to disk in advance. Instead, the black-box scorer wraps the fixed HateBERT model and exposes a query interface `predict_scores(texts)` that returns toxicity probabilities for arbitrary batches of inputs. From the perspective of the auditing algorithm, the system is accessed only via this interface.

Determinism and reproducibility notes. All random seeds are fixed for dataset processing, bias injection during fine-tuning, and auditing. The model checkpoint, label-flipping probabilities, optimizer settings, training epochs, and random seeds are logged in the experiment configuration. At audit time, inference is fully deterministic given the fixed model parameters. Together, these choices ensure reproducibility of both the ground-truth disparity and the auditing results.

A.4.2 Case Study B: Bias-in-Bios Black-Box Scoring & Reproducibility

This case study audits gender disparities in occupation prediction on Bias-in-Bios (De-Arteaga et al.,

2019). We treat a large instruction-tuned model as a black-box scorer $h^\star$ and estimate the fairness target $\Delta\mathrm{AUC}$ for a one-vs-rest task ("professor" vs. all other occupations) under limited query budgets.

Dataset. We use the HuggingFace dataset `LabHC/bias_in_bios` (splits `train`, `test`, `dev`). We concatenate splits in the fixed order `train` $\rightarrow$ `test` $\rightarrow$ `dev`, reset indices, and assign deterministic IDs `id = "ID{i}"` for $i \in \{0, \ldots, N-1\}$. We use the biography text field `hard_text`, the binary group attribute `gender` (0=male, 1=female), and the ground-truth label `profession` (integer id mapped to a string occupation name).

Occupation label set. The black-box returns a probability distribution over 28 occupations, corresponding to the columns in the cached score CSV (and exposed by `BiasInBiosBlackBox.labels`). The canonical id-to-name mapping (0..27) is: `accountant`, `architect`, `attorney`, `chiropractor`, `comedian`, `composer`, `dentist`, `dietitian`, `dj`, `filmmaker`, `interior_designer`, `journalist`, `model`, `nurse`, `painter`, `paralegal`, `pastor`, `personal_trainer`, `photographer`, `physician`, `poet`, `professor`, `psychologist`, `rapper`, `software_engineer`, `surgeon`, `teacher`, `yoga_teacher`. In the audit, we focus on the target class `professor`.

Black-box model and decoding parameters. We generate black-box scores once and cache them to disk (CSV) using the OpenAI Responses API with structured output enforcement as seen in Table 4.

Prompt specification. The system-level instructions used for scoring are:

```
You are an impartial occupation evaluation system.
Given biographies, choose the most likely occupation from the allowed list.
Return confidence scores (integers 0-100) for each allowed occupation.
Scores must sum to exactly 100.
Return ONLY valid JSON (no markdown).
Return an object with key "items" containing an array of outputs.
Return one output object per input, in the same order as inputs.
Allowed occupations: {28 labels listed above}.
```

Each output item is a JSON object with fields `id`, `occupation`, and `scores` (a dict containing all 28 label keys). The full schema is enforced via the Responses API `text.format=json_schema` with `strict=true`.

Cached score file and black-box interface. All scores are stored in a CSV with columns: `id`, `gold_occupation`, `gender`, `pred_occupation`, and 28 score columns (one per occupation). The black-box wrapper `BiasInBiosBlackBox(scores_csv)` loads this file, converts scores $s \in [0, 100]$ to probabilities $\hat{p} = s/100$, and re-normalizes row-wise so each probability vector sums to 1 (see `query_distribution`).

Fairness metric (one-vs-rest AUC for `professor`). For each biography x_i, the black-box score for the target class is $\hat{p}_i = \hat{p}(\texttt{professor} \mid x_i)$, obtained from the cached distribution. We define binary labels $Y_i = \mathbb{1}[\text{gold\_occupation}(x_i) = \texttt{professor}]$. We compute AUC separately for males and females on the audit pool: $\mathrm{AUC}_{\text{male}} = \mathrm{AUC}\big(\{\hat{p}_i, Y_i\}_{\text{gender}=0}\big)$ and $\mathrm{AUC}_{\text{female}} = \mathrm{AUC}\big(\{\hat{p}_i, Y_i\}_{\text{gender}=1}\big)$, and report the disparity $\Delta\mathrm{AUC} = \mathrm{AUC}_{\text{male}} - \mathrm{AUC}_{\text{female}}$. This $\Delta\mathrm{AUC}$ is the target quantity estimated by the active auditing pipeline in the main paper.

Determinism and reproducibility notes. All scoring uses deterministic decoding (temperature 0; top-p 1) and schema-constrained JSON outputs. Dataset IDs are deterministic given the fixed split concatenation order. The full configuration (model name, decoding parameters, label set, prompt_cache_key, truncation lengths, and CSV path) is stored alongside the cached score file and the auditing logs.

A.4.3 Hyperparameter Evaluation

Epochs for Optimization with Cooper. The number of gradient steps used in constrained optimization (`epochs_opt`) controls how accurately BAFA solves the inner C-ERM problems that produce lower and upper surrogate bounds consistent with queried black-box scores. We ablate `epochs_opt` $\in \{3, 6, 8, 10\}$ while holding λ=0.01, k=16, and `reg_alpha`=2.0 fixed, and report both query efficiency (queries to target error) and bound tightness (final width).

For BAFA-Disagreement, the `epochs_opt`=6 configuration is not reported due to missing/incomplete runs in our logs at the time of writing.

Component	Case Study A: CivilComments	Case Study B: Bias-in-Bios
Task	Hate speech / toxicity detection	Occupation inference from biographies
Dataset	CivilComments (Jigsaw Unintended Bias)	Bias-in-Bios
Audit pool size	$\sim$50k comments	$\sim$390k biographies (for comparison we take a 50k random sample)
Black-box system	Fine-tuned HateBERT classifier	OpenAI LLM via Responses API
Model identifier	`GroNLP/hateBERT`	`gpt-4.1-mini-2025-04-14`
Output signal	Toxicity probability $s_i^\star \in [0, 1]$	Integer confidence scores in $[0, 100]$
Decoding / inference	Deterministic (model in eval mode)	temperature $= 0.0$, top_p $= 1.0$
Bias mechanism	Stochastic label flipping during fine-tuning	None (natural model behavior)
Bias specification	Group-conditional flip probs (e.g. black $>$ white)	Fixed prompt + schema constraints
Fairness metric	$\Delta\mathrm{AUC} = \mathrm{AUC}_{\text{white}} - \mathrm{AUC}_{\text{black}}$	One-vs-rest ΔAUC (female vs. male)
Ground-truth disparity	$\Delta\mathrm{AUC} \approx 0.01$–$0.14$ (synthetic)	$\Delta\mathrm{AUC} \approx 0.02$–$0.05$ (observed in random sample 50k)
Caching	Not applicable (local model)	Cached once to CSV
Reproducibility	Fixed seeds, logged config	Fixed prompt, cached outputs

Table 4: Comparison of black-box setups across both case studies.

Strategy	epochs	$\epsilon = 0.02$		$\epsilon = 0.05$		Err@250	Err@$\mathbf{T}_{\max}$	Width@$\mathbf{T}_{\max}$
		Queries	Reached	Queries	Reached			
BAFA-BO	3	176 ± 132	76%	85 ± 48	97%	0.024 ± 0.015	0.025 ± 0.018	0.028 ± 0.071
BAFA-BO	6	104 ± 21	100%	53 ± 12	100%	0.019 ± 0.014	0.013 ± 0.008	0.058 ± 0.023
BAFA-BO*	8	66 ± 44	100%	47 ± 17	100%	0.018 ± 0.010	0.022 ± 0.011	0.009 ± 0.009
BAFA-BO	10	156 ± 90	91%	119 ± 52	100%	0.024 ± 0.020	0.014 ± 0.010	0.139 ± 0.160
BAFA-Dis	3	93 ± 38	56%	79 ± 35	75%	0.053 ± 0.031	0.056 ± 0.032	0.161 ± 0.452
BAFA-Dis*	8	80 ± 35	80%	64 ± 27	80%	0.017 ± 0.009	0.024 ± 0.011	0.169 ± 0.081
BAFA-Dis	10	111 ± 43	88%	78 ± 32	92%	0.021 ± 0.015	0.025 ± 0.042	0.183 ± 0.193

Table 5: **C-ERM optimization epochs ablation.** We vary `epochs_opt` (gradient steps for constrained optimization) while holding λ=0.01, k=16, and `reg_alpha`=2.0 fixed. "Queries" reports mean $\pm$ std black-box queries required to reach absolute error $\leq \epsilon$; "Reached" is the fraction of runs that reached the target within the query budget. * marks the lowest mean trajectory error configuration among those evaluated.

Batch Sizes BAFA uses two distinct batch-size parameters: the active batch size k (how many black-box queries are issued per round) and the C-ERM batch size B_{cerm} (how many queried points are processed per gradient step in Cooper). Table 6 summarises their empirical effect on the final absolute error and runtime. BAFA has two batch-size knobs: the *active* batch size k (queries per round) and the *C-ERM* batch size B_{cerm} (samples per gradient step in Cooper).

Choosing k trades off update granularity against accumulated optimisation error: smaller k triggers more frequent C-ERM solves, while larger k makes selection less responsive to changes in the certificate. Choosing B_{cerm} trades off gradient noise and stability under constraints: too small increases constraint-violation oscillations, while too large reduces the number of parameter updates per epoch for a fixed $|S_t|$ and can yield looser certificates. We found k=16 and B_{cerm}=512 to be a robust default across both case studies, providing stable C-ERM behaviour while keeping certificate updates frequent enough for effective active selection.

A.4.4 Final Case Study Hyperparameters

Can be found in Table 7.

A.4.5 Computational Costs

BAFA trades additional local computation for fewer black-box queries. Across 196 runs (828 GPU-hours total), end-to-end wall-clock time per seed is on the order of hours on a single modern

Setting	Value	Final Error
Active batch size (queries/round)		
k	8	0.0350 ± 0.0276
k	16	$\mathbf{0.0156 \pm 0.0112}$
k	32	0.0198 ± 0.0157
C-ERM batch size (samples/step)		
B_{cerm}	256	0.0232 ± 0.0137
B_{cerm}	512	$\mathbf{0.0161 \pm 0.0111}$
B_{cerm}	1024	0.0274 ± 0.0165
B_{cerm}	2056	0.0871 ± 0.0160

Table 6: **Batch size ablations (summary).** Final Error is $|\widehat{\Delta\text{AUC}} - \Delta\text{AUC}|$ at the end of the audit (mean $\pm$ std across runs).

GPU, with most time spent in the constrained optimisation step.

Hardware and runtime. Experiments ran on NVIDIA RTX A6000 (48 GB), RTX 4090, and A100 (40 GB) (single one each run). Table 8 reports wall-clock time for complete runs. CivilComments has lower per-iteration cost (2.7–5.3 min) than Bias-in-Bios (4.6–6.1 min), while the higher variance in CivilComments stems from heterogeneous hyperparameter configurations (notably `epochs_opt`) used during tuning.

Amortised cost per query. For runs targeting roughly 1200 total queries, the amortised compute cost ranges from 17–40 seconds per queried example (Table 9), with variation mainly driven by the frequency and size of C-ERM updates (smaller batches imply more optimisation rounds per fixed budget).

Where the time goes. Profiling representative runs shows that C-ERM dominates wall-clock time (about 60–70%), followed by selection (about 20–25%; BO/disagreement scoring and bookkeeping). Black-box calls contribute a smaller fraction in our local-model setting (about 5–10%) but can dominate for slow remote APIs.

Practical takeaways and speedups. Computational overhead is the main bottleneck for practitioners, but it is largely an engineering problem. The most direct improvement is to reduce how often C-ERM is solved: for example, running C-ERM every m-th iteration (or more frequently early and less frequently later) would reduce cost substantially while retaining much of the query-efficiency benefit over stratified sampling. Additional savings come from warm-starting the min/max problems from the previous round and parallelising the two C-ERM solves. In this paper we prioritise best-case query-efficiency; reducing optimisation cost is an important direction for follow-up work.

A.5 Evaluation Details

A.5.1 Evaluation Metrics

We evaluate auditing strategies using three audit-relevant metrics: convergence query-efficiency, over-time performance, and stability.

Convergence query-efficiency. Let $e_t^{(s)}$ denote the absolute estimation error after t black-box queries in run (seed) s, and let

$$\bar{e}_t := \frac{1}{S} \sum_{s=1}^{S} e_t^{(s)}$$

be the mean error across $S = 20$ seeds at query budget t. For a target accuracy threshold ε, we define the convergence query-efficiency as the smallest query budget t such that the mean error falls below the threshold,

$$t_\varepsilon := \min\{t : \bar{e}_t \le \varepsilon\}.$$

This metric reflects how many queries are required, on average across runs, to reach a desired estimation accuracy.

Over-time performance (AUEC). To capture performance throughout the auditing process, we compute the area under the error curve (AUEC) over the first $T_{\max} = 1000$ queries,

$$\text{AUEC}(T_{\max}) := \sum_{t=1}^{T_{\max}} \bar{e}_t.$$

Lower AUEC values indicate faster and more consistent error reduction over time.

Stability across seeds. To assess robustness to randomness in initialisation and sampling, we report the mean and standard deviation of the absolute error $e_t^{(s)}$ across seeds at fixed query budgets (e.g., $t = 250$). Lower variance indicates more stable auditing behaviour across runs.

A.5.2 Descriptive Statistics Results

Can be found in Table 10 and Table 11.

Parameter	CivilComments	Bias-in-Bios	Description
Experimental Setup			
Seeds	20 random seeds (0-99, sampled)		Random initialization for reproducibility
Total iterations (T)	75	75	Maximum audit rounds
Top-k batch size	16	16	Queries selected per round
Candidate pool size (M)	1000	1000	Pool size for active selection
Seed set strategy	Stratified by (g, y)		Initial labeled samples
Seed set size	$1 \times \lvert\text{groups}\rvert \times \lvert\text{labels}\rvert$		1 sample per stratum
Surrogate Model			
Architecture	`bert-base-uncased`		110M parameters, 12 layers
Max sequence length	128	128	Tokenization truncation
Learning rate	2×10^{-5}		AdamW optimizer
Batch size	16	16	Training batch size
Warmup epochs	2	2	Initial training on seed set
Retraining epochs (E_{sur})	4	4	Per-round fine-tuning
C-ERM Constrained Optimization			
Constraint tolerance (λ)	0.01	0.01	$\lvert h(x) - h^*(x)\rvert \leq \lambda$
Target precision (ϵ)	0.01	0.01	Stopping criterion (not used)
Optimization epochs (E_{opt})	10	8	Gradient steps for min/max
Optimizer batch size	512	512	Cooper constrained optimization
Regularization weight (α)	2.0	2.0	Distributional matching penalty
Optimization library	Cooper (Gallego-Posada et al., 2025)		Lagrangian-based C-ERM
Bayesian Optimization (BO strategy only)			
Acquisition function	Upper Confidence Bound (UCB)		Exploration-exploitation trade-off
UCB parameter (β)	1.0	1.0	Confidence interval width
Diversity weight (γ)	0.2	0.2	Penalty for similar queries
GP kernel	RBF (Matérn 5/2)		Gaussian Process covariance
Feature embedding	BERT [CLS] + group g		Input to GP surrogate
Black-Box Models			
Model architecture	HateBERT	GPT-4.1-mini-25-04-14	Target audited systems
Training data	SBIC (flipped labels)	Few-shot prompted	Systematic bias injection
Score range	[0, 1]	[0, 100]	Normalized to [0,1] internally
True ΔAUC	≈ 0.14	≈ 0.02–0.045	Ground-truth disparity
Datasets			
Source	CivilComments	Bias-in-Bios	Audit data pools
Task	Toxicity detection	Profession prediction	Binary classification
Protected attribute	8 identity groups	Gender (binary)	$g \in \{0, 1\}$
Pool size	$\sim$50k comments	50k random sampled biographies	After filtering
Target occupation	—	Professor vs. others	Binary task setup
Computational Resources			
GPU	RTX 4090 / A6000 / A100	RTX 4090 / A6000 / A100	24-48GB VRAM
Wall-clock time/round	$\sim$45-60s	$\sim$30-45s	Avg. over 20 seeds
Total GPU-hours/run	$\sim$4-6h	$\sim$4-6h	75 iterations

Table 7: Complete final hyperparameters for BAFA experiments across both case studies. All parameters held constant across 20 random seeds except seed initialization.

A.6 Surrogate Evaluations

A.6.1 Ablation C-ERM with smaller and larger models

An ablation over surrogate architectures (BERT-base-uncased, DistilBERT-base-uncased, and RoBERTa-base) suggests that BAFA's query efficiency is relatively insensitive to the specific surrogate choice. Figure 7 compares BAFA trajectories across three surrogate architectures. Despite substantial differences in model size and architecture, all surrogates converge to similar error levels and exhibit comparable rates of uncertainty reduction. Despite architectural differences and parameter counts ($\approx$110M for BERT-base, $\approx$66M for

Dataset	Strategy	N	Hours/run	Min/iteration
A	BAFA-BO	90	4.5 ± 4.8	2.68
	BAFA-Dis	31	6.6 ± 6.4	5.28
B	BAFA-BO	16	7.7 ± 3.6	6.14
	BAFA-Dis	17	5.7 ± 3.2	4.58

Table 8: **Runtime by dataset and strategy.** Hours/run shows mean $\pm$ std wall-clock time for complete experiments. Min/iteration is average time per audit round. The large variance in CivilComments reflects heterogeneous hyperparameter configurations across runs.

Dataset	Strategy	Total queries	Sec/query
CivilComments	BAFA-BO	800	20.1
CivilComments	BAFA-Dis	600	39.6
Bias-in-Bios	BAFA-BO	1200	23.0
Bias-in-Bios	BAFA-Dis	1200	17.2

Table 9: **Computational cost breakdown.** Sec/query is amortized cost per black-box query, including all overhead (C-ERM, BO, selection, data loading). Total GPU-h is cumulative investment across all runs.

DISTILBERT, and ≈125M for ROBERTA-base), all three surrogates reach comparable final error levels (0.0134–0.0168 at $T = 500$) and achieve $\epsilon = 0.02$ within roughly 200–350 queries in our runs. Notably, DISTILBERT converges fastest (≈200 queries), suggesting that surrogate capacity is not the primary bottleneck for audit quality in this setting (noting that this ablation uses only three seeds). This is consistent with a "version space" view of surrogate selection: the surrogate need not match the audited system's internal representations, but must approximate its input–output behavior sufficiently well to identify informative queries and keep the constraint optimization feasible. Larger autoregressive surrogates (e.g., GPT-style) may further improve alignment when auditing instruction-tuned black-box models, but this remains an empirical question and would introduce substantial compute and interface differences.

A.6.2 LoRA-surrogate Evaluation

Here, we demonstrate that the LoRA-fine-tuned BERT surrogate requires around 500 queries to mimic the black-box HateBERT model, leading us to use LoRA only for diversity embeddings, not for guiding the audit or serving as a surrogate model, as in (Yan and Zhang, 2022). This is distinct from the C-ERM surrogate, which is trained via constrained optimisation to produce the max and min bounds over the version space.

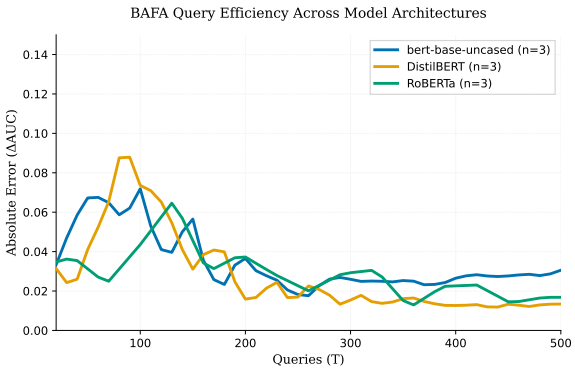

Figure 7: Reduction of uncertainty bounds for ΔAUC under different surrogate architectures. We report mean over 3 seeds.

LoRA configuration. We use Low-Rank Adaptation (LoRA) (Hu et al., 2021) to efficiently fine-tune the surrogate model in this tryout. The configuration is:

- **Base model**: BERT-base-uncased (110M parameters)
- **LoRA rank** (r): 16 (low-rank dimension)
- **LoRA alpha** (α): 32 (scaling parameter, $\alpha = 2r$)
- **LoRA dropout**: 0.1
- **Target modules**: `query` and `value` projections in attention layers
- **Trainable parameters**: ∼1.2M (1.1% of base model)

This configuration reduces memory usage by ∼90% compared to full fine-tuning while maintaining model capacity.

Surrogate metrics. We evaluate surrogate mimic behaviour using the following metrics:

- **MSE**: Mean squared error between surrogate predictions $h(x)$ and black-box scores $h^\star(x)$ on held-out data
- **Rank correlation**: Spearman/Pearson correlation of surrogate vs black-box score rankings
- **Constraint satisfaction**: Fraction of queries where $|h(x) - h^\star(x)| \leq \lambda$ (with $\lambda = 0.01$)
- **ΔAUC gap**: Difference between surrogate-computed ΔAUC and true black-box ΔAUC

Table 10: Descriptive statistics for Civil Comments dataset. For each query budget, we report mean absolute error with 95% CI, median, and IQR across all replicates.

Strategy	n	T=100		T=250		T=1000	
		Mean [95% CI]	Median (IQR)	Mean [95% CI]	Median (IQR)	Mean [95% CI]	Median (IQR)
BAFA methods							
BAFA (BO)	20	0.086 [0.066, 0.106]	0.077 (0.070)	0.021 [0.013, 0.030]	0.018 (0.020)	0.012 [0.007, 0.017]	0.013 (0.008)
BAFA (disagreement)	20	0.046 [0.028, 0.064]	0.040 (0.048)	0.020 [0.015, 0.026]	0.019 (0.017)	0.010 [0.004, 0.016]	0.007 (0.008)
Baseline methods							
BO (ablation)	20	0.067 [0.045, 0.089]	0.054 (0.065)	0.096 [0.062, 0.131]	0.088 (0.044)	0.026 [0.020, 0.033]	0.027 (0.021)
Power sampling	20	0.131 [0.092, 0.169]	0.117 (0.102)	0.108 [0.080, 0.135]	0.104 (0.089)	0.046 [0.026, 0.066]	0.030 (0.055)
Stratified sampling	20	0.095 [0.067, 0.122]	0.093 (0.079)	0.064 [0.046, 0.083]	0.064 (0.058)	0.039 [0.026, 0.052]	0.029 (0.029)

Table 11: Descriptive statistics for Bias-in-Bios dataset. For each query budget, we report mean absolute error with 95% CI, median, and IQR across all replicates.

Strategy	n	T=100		T=250		T=1000	
		Mean [95% CI]	Median (IQR)	Mean [95% CI]	Median (IQR)	Mean [95% CI]	Median (IQR)
BAFA methods							
BAFA (BO)	20	0.107 [0.058, 0.156]	0.057 (0.111)	0.022 [0.017, 0.026]	0.022 (0.011)	0.019 [0.014, 0.023]	0.016 (0.005)
BAFA (disagreement)	20	0.098 [0.051, 0.145]	0.061 (0.122)	0.022 [0.017, 0.027]	0.025 (0.016)	0.018 [0.014, 0.022]	0.019 (0.008)
Baseline methods							
BO (ablation)	20	0.043 [0.023, 0.064]	0.024 (0.050)	0.023 [0.014, 0.033]	0.013 (0.029)	0.012 [0.008, 0.016]	0.011 (0.011)
Power sampling	20	0.065 [0.035, 0.094]	0.040 (0.071)	0.065 [0.045, 0.085]	0.053 (0.064)	0.025 [0.015, 0.034]	0.021 (0.034)
Stratified sampling	20	0.058 [0.037, 0.078]	0.045 (0.065)	0.043 [0.028, 0.058]	0.036 (0.034)	0.025 [0.018, 0.033]	0.025 (0.018)

Training procedure. The surrogate is trained on the current query set S_t using a combined loss:

$$\mathcal{L} = 0.2 \cdot \mathcal{L}_{\text{MSE}} + 0.8 \cdot \mathcal{L}_{\text{rank}},$$

where $\mathcal{L}_{\text{MSE}}$ is mean squared error between surrogate probabilities and black-box scores, and $\mathcal{L}_{\text{rank}}$ is a margin ranking loss that preserves pairwise score orderings. Training uses AdamW optimizer with learning rate 5×10^{-4}, batch size 16, and 4 epochs per iteration.

Surrogate–black-box agreement and implications for constraint-based auditing. Figure 8 demonstrates how quick a BERT-based LoRA-surrogate (results are very similar with HateBERT as a surrogate) approaches the audited system as the query budget grows. Pointwise score agreement and rank correlation increase steadily and reach high values after a few hundred queries, but accurately reproducing the audit target requires substantially more supervision. In both settings (BERT and HateBERT), the surrogate's induced disparity estimate $\Delta\text{AUC}(h)$ aligns quantitatively with the black-box disparity $\Delta\text{AUC}(h^\star)$ only after roughly 500–750 queried examples. Before this query intervall, the surrogate often captures the correct direction of the disparity but exhibits large magnitude error in $\left|\Delta\text{AUC}(h^\star) - \Delta\text{AUC}(h)\right|$, indicating that matching scores in an average sense is not sufficient to match the groupwise ranking geometry that determines ΔAUC.

This gap matters for approaches that impose surrogate-based constraints in C-ERM, e.g., methods in the spirit of (Yan and Zhang, 2022) that treat $h(x)$ as a proxy for $h^\star(x)$ inside the constraint set. In our setting, reaching the regime where surrogate-based constraints would be reliable already consumes a significant fraction of the overall query budget, weakening the case for query-efficient third-party auditing. We therefore avoid using a learned surrogate as a constraint proxy in the certificate computation: BAFA's certificate interval is derived by constrained optimisation using queried black-box scores only, not surrogate predictions as (Yan and Zhang, 2022) are doing. Learned representations are used only as auxiliary signals in the selection module (e.g., diversity-aware selection and BO features). Finally, to reflect realistic audit conditions for ranking-based metrics such as ΔAUC, we adopt a top-k selection procedure that leverages known ground-truth labels for evaluation, rather than relying on surrogate-imputed scores. Together, these design choices keep BAFA effective in the low- to mid-budget regime where surrogate-based constraints are not yet dependable as visibly.

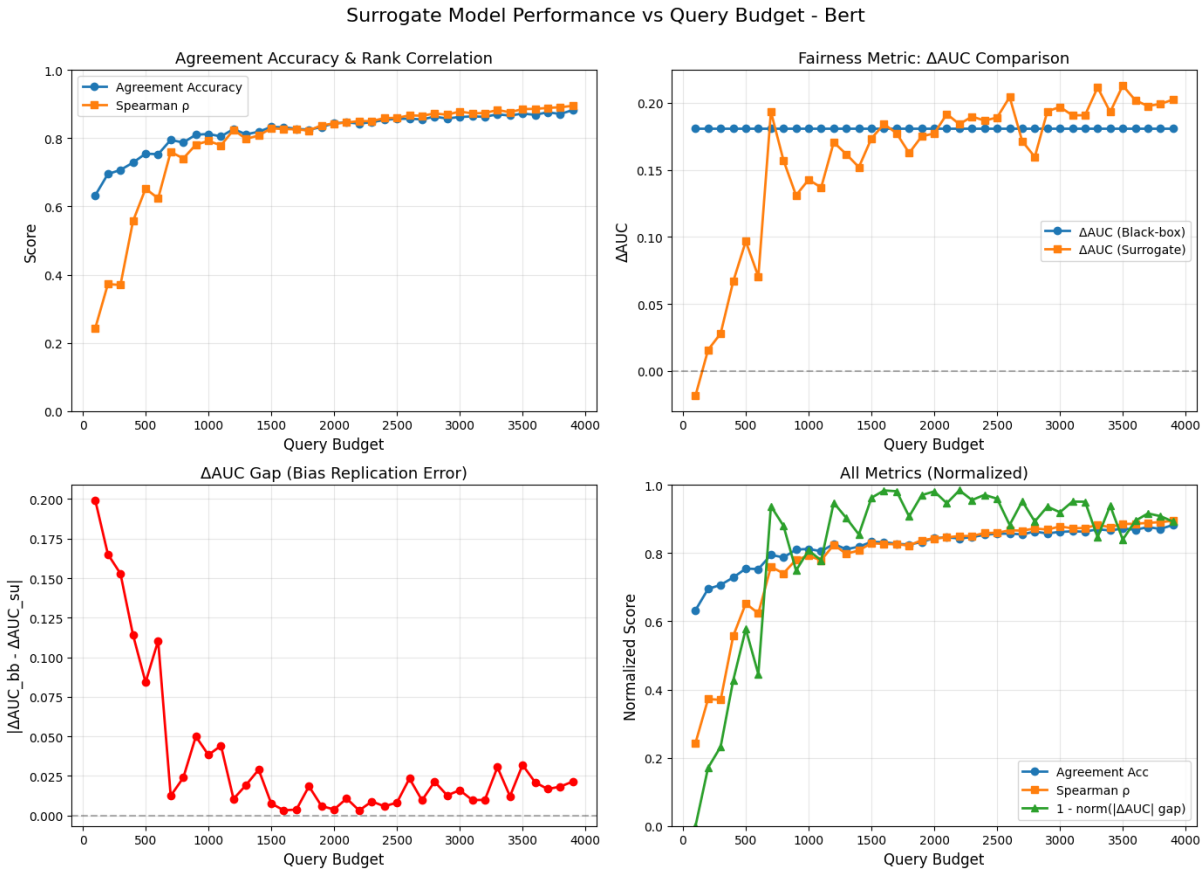

Figure 8: **Surrogate–black-box agreement vs. query budget.** As the queried set grows, we report (i) pointwise agreement accuracy and Spearman rank correlation between surrogate scores $h(x)$ and black-box scores $h^\star(x)$, and (ii) the induced disparity replication error $\left|\Delta\mathrm{AUC}(h^\star) - \Delta\mathrm{AUC}(h)\right|$. While agreement and rank correlation increase steadily, the ΔAUC replication error becomes small and stable only after roughly 500–750 queries, indicating that many queries are required before the surrogate matches the black-box score geometry relevant for groupwise ranking.