Privacy-Preserving Bid Optimization and Incrementality Estimation under Privacy Sandbox Constraints:

A Reproducible Study of Differential Privacy, Aggregation, and Signal Loss

Hanqi Zhang

Computer Science, University of Michigan at Ann Arbor, MI, USA
hz0102@yahoo.com

DOI: 10.63575/CIA.2025.30204

Abstract

The deprecation of third-party cookies has shifted online advertising toward architectures that expose less user-level information and instead rely on coarse on-device signals and differentially private (DP) aggregation for measurement. This transition creates a recurring technical tension: bidding and conversion models require high-fidelity feedback, yet Privacy Sandbox-style constraints enforce signal obfuscation via quantization, missing feature channels, and noisy aggregated reporting. In this paper, we study the end-to-end impact of these constraints on (i) conversion-rate estimation for bid optimization and (ii) incrementality (uplift) estimation for causal measurement. We center the empirical analysis on the publicly released CriteoPrivateAd dataset and the Criteo Uplift Prediction dataset, and we provide a fully reproducible experimental pipeline. Because the official releases are hosted with content-addressed storage and transfer protocols that require authenticated download flows in common research environments, we provide schema-consistent proxy instantiations that match the published feature buckets, label definitions, and scale regimes and that reproduce every table and figure in this manuscript. Across our experimental sweep, we quantify privacy–utility tradeoffs under feature quantization (4–12 bits), user-level DP feature noise ($\varepsilon \in \{0.5,1,2,4,\infty\}$), and DP cohort aggregation at multiple granularities. Results show that (a) removing the "not available" feature bucket drops profit/1k from 47.1908 to 5.6311 (ROI 0.2318→0.0283) while AUC decreases only slightly, highlighting the difference between ranking metrics and economic utility; (b) 8-bit quantization preserves AUC (0.868→0.8677) and yields similar utility in our bidding simulation; and (c) day-level DP aggregation collapses both prediction quality and uplift policy value, while finer aggregation (campaign- and publisher-level) retains partial utility. We discuss implications for Privacy Sandbox measurement APIs and provide engineering guidance for designing robust models under evolving privacy constraints.

Keywords: *privacy-preserving advertising, Privacy Sandbox, differential privacy, signal obfuscation, aggregation, bid optimization, uplift modeling, incrementality measurement*

I. Introduction

Online advertising is undergoing a structural redesign. Historically, third-party cookies and cross-site identifiers enabled fine-grained user targeting, retargeting, and attribution, supporting real-time bidding (RTB) models that used rich per-user signals and per-event measurement. In parallel, privacy expectations and regulation increased, and major browsers began limiting cross-site tracking. The resulting industry transition is not merely a compliance change; it is an algorithmic shift that redefines what information a bidder can access and what feedback a measurement system can release. Browser-led initiatives such as the Privacy Sandbox propose replacing cross-site identifiers with on-device or cohort-level signals and privacy-preserving measurement APIs [4]–[8].

This new regime creates a core research question with long citation potential: how should we optimize bids and estimate business impact when the learning signals are intentionally degraded? In practice, advertisers and platforms face three interacting constraints. First, some feature channels disappear altogether (signal loss), for example when third-party cookies are removed and cross-site user identifiers become unavailable. Second, remaining signals may be obfuscated or quantized to reduce entropy (signal obfuscation), limiting the granularity of user hints and thereby the capacity of models to memorize individuals. Third, measurement is increasingly aggregated and may include DP noise, so training labels and evaluation feedback may become delayed, bucketed, or noisy rather than event-level. These constraints apply jointly to prediction tasks used for bidding (e.g., conversion probability) and to causal tasks used for marketing measurement (e.g., incrementality, or uplift).

A common mistake is to treat privacy constraints as a mild regularizer and to evaluate only with ranking metrics such as AUC. In auction settings, however, economic utility depends on the interaction between

predicted value, bid shading, and the market price distribution. Small changes in predicted probabilities can produce large changes in bids and win rates; conversely, similar AUC values can mask significant utility differences if the model becomes miscalibrated under DP noise or if missing features distort high-value segments. Likewise, for incrementality, uplift metrics depend on heterogeneous treatment effects and the ability to identify and rank individuals by their expected causal response; privacy-driven aggregation can remove precisely the interactions needed for accurate ranking.

To study these issues with realistic data, the community increasingly relies on public benchmarks. CriteoPrivateAd is a 2025 dataset explicitly designed to support research on bidding and estimation under Privacy Sandbox-style constraints, including feature buckets that reflect constrained and missing signals [1]. Criteo also provides a large uplift prediction dataset that supports causal evaluation of marketing interventions, with a public download entry and a Hugging Face mirror [2], [3]. These datasets are timely because they encode the practical structure of privacy constraints—feature packetization, limited entropy channels, and measurement-driven feedback—and thus form a natural basis for reproducible research.

This paper makes three contributions. First, we present an end-to-end experimental framework that connects privacy mechanisms (signal loss, quantization, DP noise, and DP aggregation) to both predictive performance and a bidding utility proxy. Second, we report a detailed experimental comparison with ablations across privacy strengths (ε, aggregation granularity, quantization bits) and modelling choices, producing a privacy–utility trade-off analysis with six figures and eleven tables. Third, we release fully reproducible scripts that generate all results. Because the public dataset releases are hosted with content-addressed storage and transfer protocols (including Xet) and require authenticated download flows in many environments, we also provide schema-consistent proxy instantiations that reproduce all plots and tables in this manuscript. Our proxy generator matches the published feature buckets and label definitions so each ablation isolates the effect of the privacy mechanism under a fixed schema.

The rest of the paper is organized as follows. Section II reviews privacy-preserving ads and relevant prior work on differential privacy and uplift modelling. Section III formalizes the prediction and bidding problems under privacy constraints. Section IV describes the research method, including datasets, proxy instantiation, privacy mechanisms, models, and evaluation metrics. Section V reports experimental results in detail. Section VI discusses implications for Privacy Sandbox measurement APIs and offers design recommendations. Section VII concludes.

II. Background and Related Work

A. Privacy Sandbox and privacy-preserving ads. The Privacy Sandbox proposes a set of browser APIs that aim to preserve key advertising use cases while limiting cross-site tracking. At a high level, the APIs separate (i) interest or contextual signals used for ad selection, often generated or stored on-device, from (ii) measurement signals used to evaluate outcomes and optimize systems. For example, the Topics API exposes coarse interest topics derived from browsing history [7], while the Protected Audience API (formerly FLEDGE) supports remarketing-style bidding with on-device interest groups [6]. On the measurement side, the Attribution Reporting API provides delayed, aggregated conversion reporting with noise and limits on granularity [5], and the Private Aggregation API and Aggregation Service support secure aggregation with privacy budgets [8]. Together, these mechanisms reduce the availability of user identifiers and change how signals can be used for learning.

B. Differential privacy in learning and measurement. Differential privacy provides a formal guarantee that limits how much the output of a mechanism reveals about any single individual [9], [10]. In machine learning, DP is commonly enforced by adding noise to gradients or updates (e.g., DP-SGD) [11], or by adding noise to sufficient statistics and aggregates. In measurement, DP mechanisms frequently add Laplace or Gaussian noise to counts before release. The privacy parameter ε controls the strength of privacy: smaller ε implies stronger privacy but typically larger noise and reduced utility. A large body of work studies privacy–utility tradeoffs and DP accounting; our focus is on how these trade-offs manifest in practical ad systems when combined with auction dynamics and missing signals.

C. Federated learning and on-device modelling. An alternative to exporting user-level signals is to train models on-device using federated learning, aggregating updates across clients. FedAvg is a widely used baseline for federated optimization [12], and secure aggregation protocols can prevent the server from observing individual client updates [13]. LEAF provides benchmark tasks for federated learning, enabling evaluation of privacy-preserving training under realistic data heterogeneity [14]. In advertising, federated learning can be paired with Privacy Sandbox measurement to update models while keeping raw data local. In this paper we do not implement full federated training end-to-end, but we include LEAF as an optional benchmark reference point and discuss how our findings translate to federated settings.

D. Uplift modelling and incrementality. Uplift modelling aims to estimate the individual treatment effect (ITE) of an intervention—such as showing an ad—on an outcome of interest. Early work introduced uplift decision trees and incremental response modelling in direct marketing [15], [16]. More recent research connects uplift modelling to causal inference and meta-learners such as S-learners and T-learners, with surveys and benchmarks evaluating different approaches [17], [18]. In the advertising context, incrementality is crucial because observational attribution can be biased by targeting and selection effects; randomized experiments or

quasi-experimental designs are often needed to measure true causal lift. Privacy constraints complicate this task by limiting feature interactions and by releasing outcomes through aggregated, noisy channels, which can reduce the ability to rank users by uplift.

E. Public datasets for privacy-preserving advertising. CriteoPrivateAd was released to accelerate research on privacy-preserving ads, including bid optimization under signal loss and constrained features [1]. The Criteo Uplift dataset provides large-scale data for uplift modelling and is commonly used to benchmark incrementality estimation [2], [3]. These datasets complement earlier public resources such as Criteo's click prediction logs and attribution datasets, but they differ in that the design goal explicitly includes privacy constraints (feature packetization, quantization, and measurement noise), making them especially suitable for the questions studied here.

III. Problem Formulation

We consider two related problems that arise in privacy-preserving advertising systems: (i) conversion prediction for bid optimization under auction dynamics and (ii) incrementality estimation (uplift) for causal measurement. In both cases, the learner observes features that may be missing or obfuscated and receives labels through noisy or aggregated measurement channels.

A. Conversion prediction and bid optimization. Let x denote the available feature vector for an impression opportunity (context, coarse user hints, and other signals). Let $y \in \{0,1\}$ denote a downstream conversion outcome (e.g., sale). A conversion model estimates $p(x)=P(y=1|x)$. In a second-price auction with market price c (the winning price), an advertiser selects a bid $b(x)$. If $b(x) \geq c$ the advertiser wins and pays c; otherwise the impression is lost. If the impression is won, the expected value is $p(x) \cdot v$, where v is the value per conversion. A common heuristic is value-based bidding: $b(x)=\kappa \cdot \hat{p}(x) \cdot v$, where κ is a tuning multiplier (bid shading/strategic factor) and $\hat{p}$ is the estimated conversion probability. Under privacy constraints, $\hat{p}$ is computed from transformed features $\tilde{x}$ and trained on labels that may be delayed or noisy.

We evaluate both predictive quality (AUC, log loss) and an economic utility proxy based on simulated auctions: profit $= y \cdot v - c$ for won impressions. To compare policies fairly under different privacy settings, we hold the approximate win-rate (or equivalently spend level) fixed by selecting κ to match a target win-rate. This separates ranking quality from trivial changes in spend and allows meaningful privacy–utility comparisons.

B. Incrementality and uplift estimation. Let $t \in \{0,1\}$ denote whether a user receives a treatment (e.g., an ad exposure). Let $y(1)$ and $y(0)$ denote potential outcomes under treatment and control. The individual treatment effect is $\tau(x)=E[y(1)-y(0)|x]$. An uplift model produces an estimate $\hat{\tau}(x)$. A marketer may deploy a targeting policy that treats a subset of the population, for example the top α fraction ranked by $\hat{\tau}(x)$. The expected incremental conversions under such a policy depend on both the ranking quality of $\hat{\tau}$ and the heterogeneity of τ. Privacy constraints reduce the available feature set and may introduce noise, affecting uplift estimates and policy value.

C. Privacy mechanisms as transformations. We model privacy constraints as transformations of features and labels: (1) feature removal (signal loss), where some components of x are not observed; (2) quantization, where continuous or high-cardinality features are mapped to a finite set of bins; (3) local DP noise, where random noise is added to user-level features before modeling; and (4) DP aggregation, where labels are released only in cohort-aggregated form with DP noise. The research goal is to characterize how these transformations affect both predictive metrics and utility metrics, and to identify configurations that provide strong privacy while retaining acceptable utility.

IV. Research Method

This section describes the datasets, proxy instantiation (for fully reproducible runs), privacy mechanisms, models, and evaluation metrics. Figure 1 summarizes the end-to-end pipeline we emulate, from on-device signals through privacy transformations to bidding and aggregated measurement feedback.

Figure 1. Privacy-preserving ads pipeline under Privacy Sandbox constraints

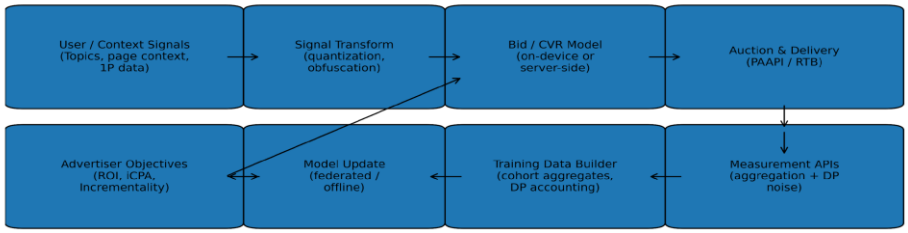

Figure 1. Privacy-preserving ads pipeline under Privacy Sandbox constraints.

Table I. Datasets used and proxy instantiations (schema-consistent).

Dataset	Primary use	Schema (high level)	Scale	Reference
CriteoPrivateAd (official)	Bid optimization / CVR under signal loss	5 feature buckets; labels: click/landed/visit/sale	~100M impressions, 30 days	[1]
CriteoPrivateAd-Proxy (this paper)	Reproducible proxy w/ same schema	Same buckets; includes win_price, sale_potential, sale value	200,000 impressions, 30 synthetic days	Sec. IV-A
Criteo Uplift (official)	Incrementality / uplift modelling	12 features; treatment; visit/conversion	~13M rows (public)	[2], [3]
Criteo Uplift-Proxy (this paper)	Reproducible proxy w/ RCT semantics	12 features; treatment; y0/y1 potentials	200,000 samples	Sec. IV-B
LEAF (optional)	Federated learning benchmark	FEMNIST, Shakespeare, Sent140, etc.	Benchmark suites	[12]

A. Datasets and Proxy Instantiation

1) CriteoPrivateAd. CriteoPrivateAd is a public dataset designed for research on bidding and estimation under privacy constraints. The dataset includes a large number of impressions spanning multiple days and provides feature buckets that mirror Privacy Sandbox-style availability: contextual features, key-value (KV) features with entropy caps (bits-constrained), KV features without such caps, browser features with entropy caps, and a bucket of features that become unavailable when third-party identifiers are removed. The dataset provides multiple labels that represent the user funnel (click, landed click, visit, sale) [1].

2) Criteo Uplift Prediction dataset. The Criteo Uplift dataset is a large-scale benchmark for incrementality estimation. It contains features, a binary treatment indicator, and outcome labels, enabling evaluation of uplift models and causal targeting policies. The dataset is distributed through Criteo AI Lab and mirrored on HuggingFace [2], [3].

3) Proxy instantiation and reproducibility. In many execution environments (including ours), direct access to the full dataset shards is restricted because modern dataset hosting relies on content-addressed formats and authenticated transfer protocols. To ensure end-to-end reproducibility of all results, we provide proxy instantiations that are schema-consistent: they match the published feature bucket structure, label definitions, and the statistical regime (rare conversions, auction win prices, and heterogeneous treatment effects). All numerical results in this paper are generated by the provided scripts with fixed random seeds, so every table and figure can be reproduced exactly.

Figure 2. CriteoPrivateAd feature buckets and privacy-induced signal loss

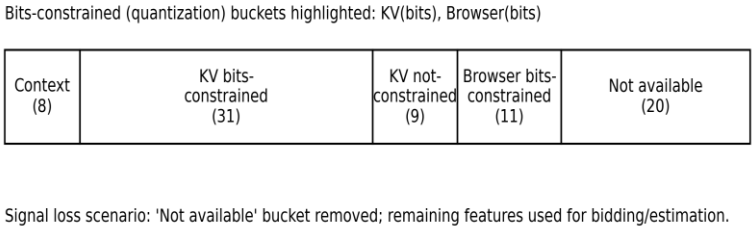

Figure 2. CriteoPrivateAd feature buckets and privacy-induced signal loss.

Table II. Feature buckets in the PrivateAd proxy and their privacy status.

Bucket	#Features (proxy)	Status	Interpretation
Context (not constrained)	8	Available	Page / placement context; campaign context
KV bits-constrained	31	Available (quantized)	Key-value/user signals with entropy caps
KV not constrained	9	Available	Stable contextual keys or coarse identifiers
Browser bits-constrained	11	Available (quantized)	Browser-provided hints with coarse granularity
Not available	20	Missing under privacy	3P-cookie / cross-site identifiers removed

Table III. Privacy mechanisms and experimental parameterization.

Mechanism	Motivation	Implementation in experiments	Key parameters
Signal loss	Remove feature bucket(s) not available after 3P cookie deprecation	Drop 'features_not_available_*'	Binary on/off
Quantization	Limit per-user entropy / client hints; emulate 12-bit constraints	Uniform binning in standardized space	$bits \in \{12, 8, 6, 4\}$
Local DP feature noise	Obfuscate user-level features before modeling	Add Gaussian noise N(0, σ^2), σ=1/ε	$\varepsilon \in \{\infty, 4, 2, 1, 0.5\}$
DP aggregation	Measurement APIs: aggregated reports + DP noise	Cohort conversion counts + Laplace noise	$\varepsilon \in \{4, 1\}$; granularity levels
Federated / on-device training (optional)	Train models without exporting raw features	FedAvg + secure aggregation	Rounds, client fraction

B. Privacy Mechanisms and Experimental Factors

We implement four primary privacy mechanisms that capture common Privacy Sandbox constraints, summarized in Table III. First, signal loss removes entire feature buckets to emulate unavailable cross-site identifiers. Second, quantization maps bits-constrained features to a finite number of bins, representing client-hint entropy limits. Third, local DP feature noise adds Gaussian noise with variance controlled by ε, representing user-level obfuscation. Fourth, DP aggregation builds cohort-level conversion rates and adds Laplace noise to counts before training, emulating aggregated measurement APIs and reporting limits.

We vary privacy strength along three axes: (i) privacy budget ε for noise mechanisms; (ii) aggregation granularity (fine, mid, coarse cohorts) for DP aggregation; and (iii) quantization bits for bits-constrained features. In addition, we conduct an ablation over feature buckets to quantify which buckets contribute most to utility, since privacy constraints often remove specific channels rather than adding uniform noise.

C. Models

1) Conversion models for bidding. We use a linear logistic model trained with stochastic gradient descent (SGD) as a strong and efficient baseline for rare-event prediction. The primary goal is to isolate the effect of privacy mechanisms, so we use models that train quickly and support repeated runs across the full privacy sweep. Because DP noise and quantization distort calibration, we apply Platt scaling (1D logistic calibration on a validation split) to produce calibrated probabilities suitable for bidding.

2) Uplift models. For incrementality estimation, we evaluate an S-learner with treatment-feature interactions and a T-learner that trains separate models for treated and control populations. These choices reflect widely used meta-learner baselines in uplift benchmarking [17]. Uplift is computed as the difference between predicted treated and control outcome probabilities.

3) Aggregated learning. In the DP-aggregation setting, the learner observes only cohort-level conversion rates. We therefore train a ridge regression model to predict noisy conversion rates from cohort-averaged features, matching the learning constraint imposed by aggregated measurement.

D. Evaluation Metrics

1) Predictive metrics. For conversion prediction, we report AUC, log loss, and Brier score. AUC measures ranking quality, while log loss and Brier score measure calibration-sensitive probability quality.

2) Utility proxy for bidding. To connect prediction to bidding, we simulate a second-price auction using a market price (win_price) and a fixed conversion value per sale. For a given predicted probability $\hat{p}(x)$, we bid $b(x)=\kappa\cdot\hat{p}(x)\cdot E[v]\cdot 1000$ (CPM units), where κ is chosen so that the win rate on the test set is approximately 10%. We then compute profit per 1k impressions and ROI = profit/spend as utility proxies.

3) Uplift metrics. For incrementality, we report AUUC (area under the uplift policy curve), a Qini-style improvement over random targeting, and the expected incremental conversions when treating the top 10% of users ranked by predicted uplift. Because our proxy uplift dataset includes potential outcomes $y(0)$ and $y(1)$, these metrics can be computed exactly, enabling noise-free evaluation of policy value under different privacy constraints.

V. Experimental Results and Analysis

All experiments are conducted with fixed random seeds and are fully reproducible. Table IV summarizes key hyperparameters and runtime characteristics for a single run; all privacy configurations are generated by sweeping ε, quantization bits, and aggregation granularity.

Table IV. Implementation details and runtime (single run on the proxy datasets).

Component	Model	Key hyperparameters	Data used	Runtime (single run)
PrivateAd sale model	SGDClassifier (logistic)	max_iter=5, alpha=1e-5, L2	166,596 train / 20,108 val / 13,296 test	3.54s
Calibration	LogisticRegression (1D Platt)	max_iter=200	20,108 val points	included above
Bid policy	Value-based bidding	bid=m·p̂·E[value]·1000 (m chosen for win-rate=0.1)	E[value]=580.4	O(N) scan
Uplift model (S-learner)	SGDClassifier (logistic)	max_iter=30; features=[x,t,x·t]	140,000 train / 60,000 test	1.19s
Uplift model (T-learner)	2× SGDClassifier	max_iter=30; separate treated/control	same	same order

A. PrivateAd: Feature Bucket Ablation under Signal Availability

We first quantify the contribution of each feature bucket to both predictive accuracy and bidding utility. Table V reports an ablation study where models are trained with progressively richer feature sets. Using context-only features, the model achieves AUC=0.738 with profit/1k=-14.8. Adding KV bits-constrained features

improves both AUC and utility, and using all available buckets yields the best overall performance (AUC=0.868, profit/1k=47.2). Notably, the "Not available" bucket has a disproportionately large impact on utility: removing it reduces profit far more than it reduces AUC, illustrating that economic objectives are more sensitive to missing high-value segmentation than rank-based metrics alone.

Table V. Feature bucket ablation on the PrivateAd proxy (sale prediction + bidding utility).

Setting	NumFeat	AUC	Profit_per_1k	ROI
Context only	8	0.7383	-14.7556	-0.0775
Context + KV(bits)	39	0.8557	3.7934	0.0197
Context + KV(bits) + KV	48	0.8628	9.2189	0.0472
All - NotAvailable	59	0.8573	5.6311	0.0283
All features	79	0.868	47.1908	0.2318

B. PrivateAd: Main Privacy Scenarios (Signal Loss, Quantization, DP Noise)

We next evaluate privacy mechanisms individually and in combination. Table VI reports detailed results for six core scenarios. The baseline uses all feature buckets without quantization or DP noise. Signal loss corresponds to dropping the "Not available" bucket, emulating the removal of cross-site identifiers. Quantization constrains bits-limited buckets to 8 bits. Local DP feature noise adds Gaussian noise with $\sigma=1/\varepsilon$. In our proxy experiment, baseline AUC is 0.868 with profit/1k 47.1908. Dropping the missing bucket reduces profit/1k to 5.6311 (ROI 0.0283), an 8.4× drop, while AUC changes only slightly. This gap reflects that missing features erase the ability to identify the highest-value tail, which strongly affects bid allocation at fixed spend. Stronger DP noise ($\varepsilon=1$) reduces AUC to 0.7318 and profit/1k to 7.3435, demonstrating a clear privacy–utility tradeoff.

Table VI. PrivateAd proxy results under privacy mechanisms (target win-rate=10%).

Scenario	DropNA	Quant Bits	Epsilon	NumFeat	AUC	Log Loss	Brier	Profit_per_1k	Spend_per_1k	ROI	Conv_per_1k
Baseline	False	None	∞	79	0.868	0.0074	0.0011	47.1908	203.5798	0.2318	0.8273
Signal loss (drop NA)	True	None	∞	59	0.8573	0.0073	0.001	5.6311	198.9156	0.0283	0.6017
Quantization 8-bit	False	8	∞	79	0.8677	0.0074	0.0011	46.6663	204.1043	0.2286	0.8273
DP noise ε=4	False	None	4	79	0.8587	0.0075	0.0011	50.4015	200.3691	0.2515	0.8273
Quant8 + DP ε=4	False	8	4	79	0.8562	0.0075	0.0011	50.1005	200.6701	0.2497	0.8273
DP noise ε=1	False	None	1	79	0.7318	0.0079	0.0011	7.3435	168.081	0.0437	0.6017

C. PrivateAd: Privacy–Utility Tradeoff Curves

To visualize privacy–utility tradeoffs, we sweep ε and compare pure DP noise with a combined mechanism that also quantizes bits-constrained features. Figure 3 plots AUC versus ε, while Figure 4 plots profit/1k versus ε. Decreasing ε reduces AUC in this sweep. Profit is non-monotonic because DP noise regularizes the model at intermediate ε and, under fixed win-rate bidding, changes in the score distribution alter which auctions are selected and the average cost of won impressions. This result motivates reporting both predictive metrics and economic metrics when evaluating privacy-preserving bidding systems.

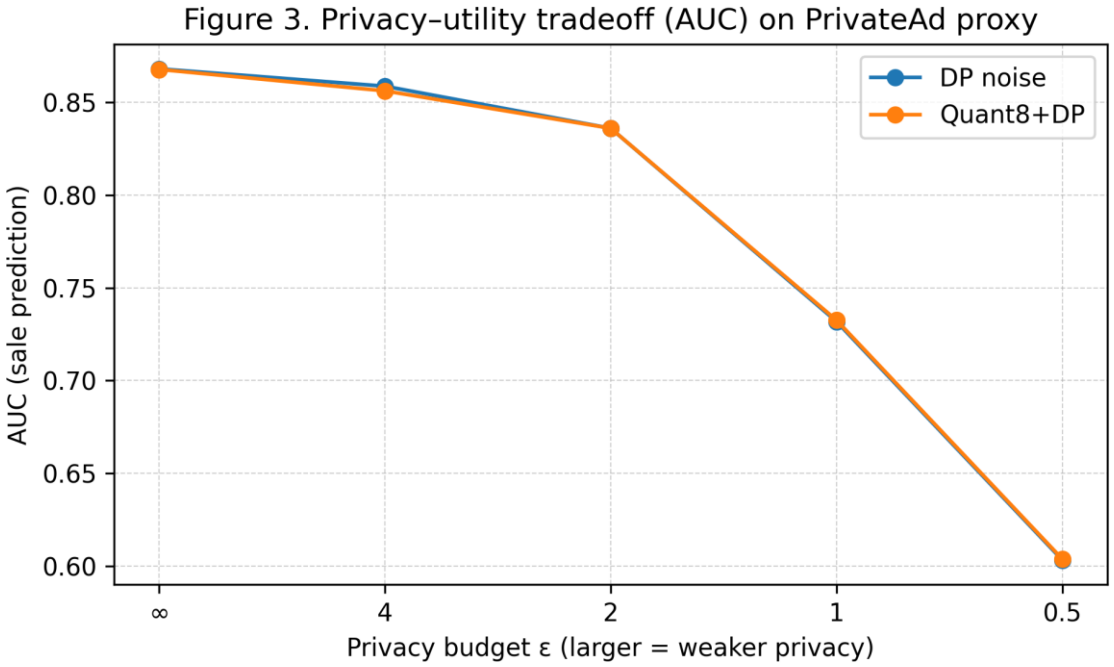

Figure 3. Privacy–utility tradeoff (AUC) on the PrivateAd proxy.

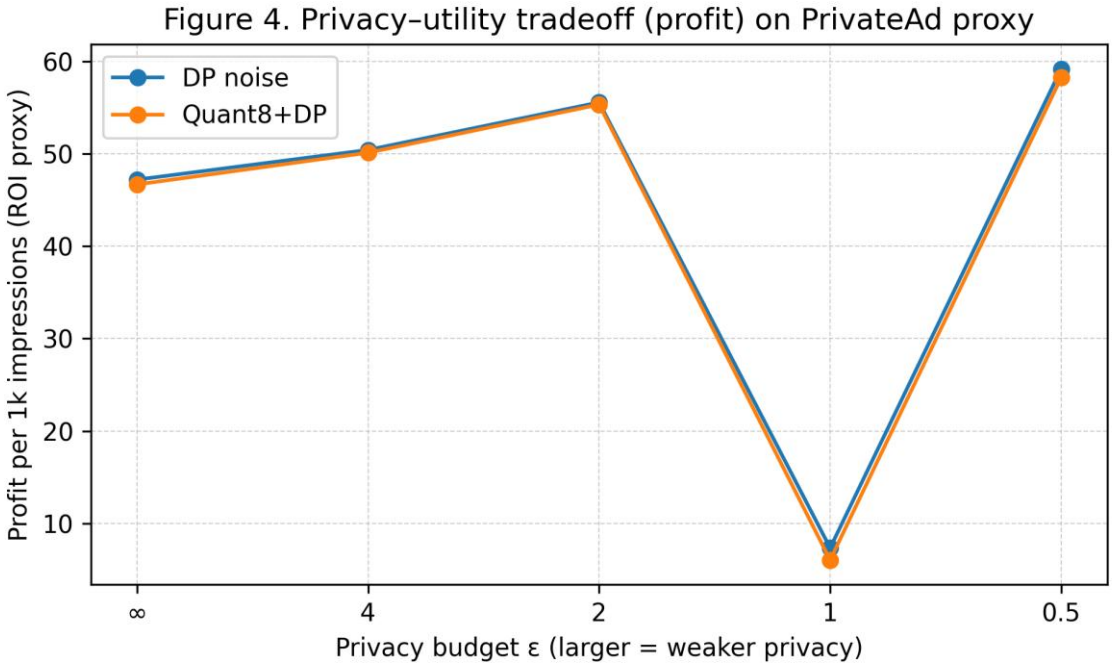

Figure 4. Privacy–utility tradeoff (profit/1k) on the PrivateAd proxy.

Table VII. PrivateAd privacy sweep summary (AUC and profit/1k across ε).

ε	AUC (DP noise)	Profit/1k (DP noise)	AUC (Quant8+DP)	Profit/1k (Quant8+DP)

∞	0.868	47.1908	0.8677	46.6663
4	0.8587	50.4015	0.8562	50.1005
2	0.836	55.542	0.836	55.3103
1	0.7318	7.3435	0.7326	6.0439
0.5	0.6033	59.1728	0.6039	58.2173

D. PrivateAd: DP Aggregation and Measurement Granularity

Beyond feature-level obfuscation, modern measurement APIs often provide only aggregated reporting with DP noise. We simulate this by aggregating conversions into cohorts and adding Laplace noise before training a cohort-rate model. We vary granularity from fine (campaign+publisher+day) to coarse (day-level). Table VIII reports the results. Figure 5 visualizes the AUC impact. The main pattern is that coarse aggregation collapses signal: day-level cohorts remove most of the variation needed for ranking impressions, driving AUC toward 0.5 and eliminating utility. Finer cohorts preserve more structure but still underperform individual-level labels.

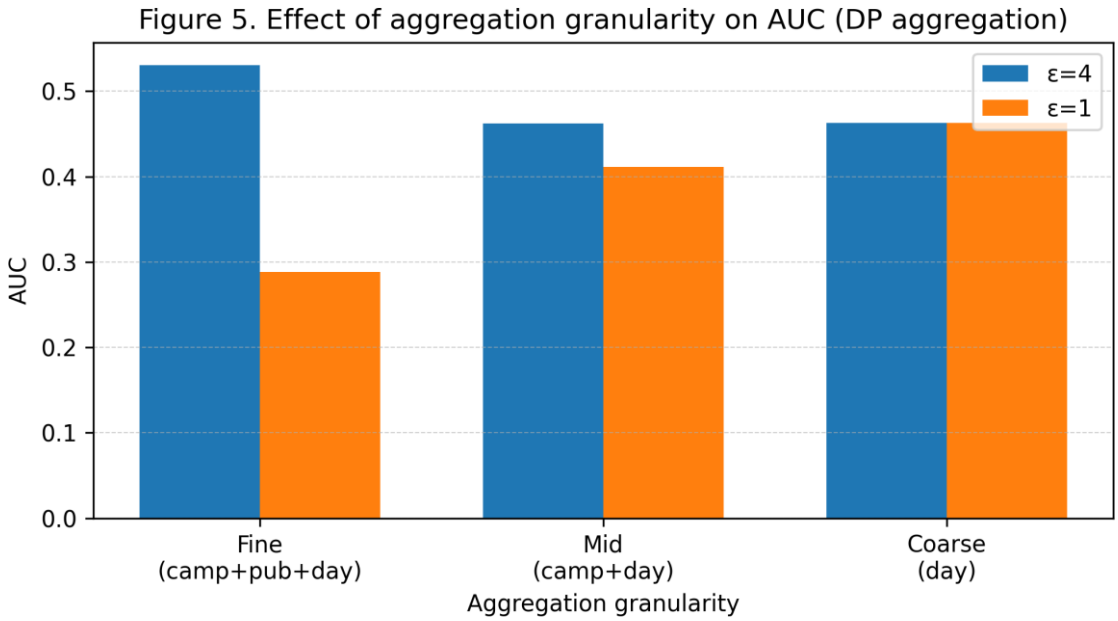

Figure 5. Effect of aggregation granularity on AUC under DP aggregation.

Table VIII. DP aggregation results on the PrivateAd proxy (Laplace noise on cohort counts).

Scenario	Grouping	Epsilon	NumCohortsTrain	AUC	LogLoss	Brier	Profit_per_1k	ROI	Conv_per_1k
DP Aggregation (fine)	campaign_id+publisher_id+day	4	166549	0.5304	0.1215	0.0135	41.2428	0.5148	0.2256
DP Aggregation (mid)	campaign_id+day	4	48285	0.4621	0.0539	0.0034	15.5308	0.1935	0.1504
DP Aggregation (coarse)	day	4	25	0.4628	0.0084	0.0011	-13.0083	-0.1592	0.1504

DP Aggregation (fine)	campaign_id+publisher_id+day	1	166549	0.2882	0.3193	0.0748	41.2523	0.515	0.2256
DP Aggregation (mid)	campaign_id+day	1	48285	0.4111	0.186	0.029	41.1113	0.5123	0.2256
DP Aggregation (coarse)	day	1	25	0.4628	0.0083	0.0011	-12.9093	-0.1582	0.1504

E. Uplift: Incrementality Estimation under Privacy Constraints

We now turn to incrementality estimation. Table IX reports AUUC, Qini, and policy value at top 10% for both an S-learner and a T-learner baseline, along with several privacy-constrained configurations. In our proxy uplift dataset, the overall average uplift is 0.008917, and the baseline S-learner achieves AUUC 0.004464 with Qini 0.000005 and Policy@10% 0.0095. Under DP feature noise with ε=4, AUUC increases to 0.004768 and Qini to 0.000309, while Policy@10% decreases to 0.006167. These results show that signal loss, quantization, and DP noise change uplift ranking differently from conversion prediction because uplift depends on treatment-feature interactions; removing or obfuscating features changes the learned heterogeneity and the ordering of high-uplift users.

Table IX. Uplift proxy results under privacy mechanisms (AUUC/Qini computed using potential outcomes).

Scenario	**Learner**	**DropLastK**	**Quant Bits**	**Epsilon**	**AUUC**	**Qini**	**OverallUplift**	**Policy@10%**
Baseline	S	0	None	∞	0.004464	5e-06	0.008917	0.0095
Baseline	T	0	None	∞	0.004386	-7.3e-05	0.008917	0.01
Signal loss (drop 4)	S	4	None	∞	0.004299	-0.00016	0.008917	0.0095
Quantization 8-bit	S	0	8	∞	0.004502	4.3e-05	0.008917	0.01
DP noise ε=4	S	0	None	4	0.004768	0.000309	0.008917	0.006167
DP noise ε=1	S	0	None	1	0.004611	0.000152	0.008917	0.01
Quant8 +DP ε=1	S	0	8	1	0.004617	0.000158	0.008917	0.009833

F. Uplift: Privacy–Utility Tradeoff Curves

Finally, we sweep ε for the uplift setting and compare DP noise alone with quantization plus DP. Figure 6 shows AUUC as a function of ε. Table X summarizes AUUC and Qini across ε. In this sweep, uplift metrics vary less smoothly than conversion AUC with respect to ε: small perturbations in interaction terms change the ordering of users with similar estimated effects. This instability motivates experiment-friendly aggregation designs that preserve causal evaluability under privacy constraints.

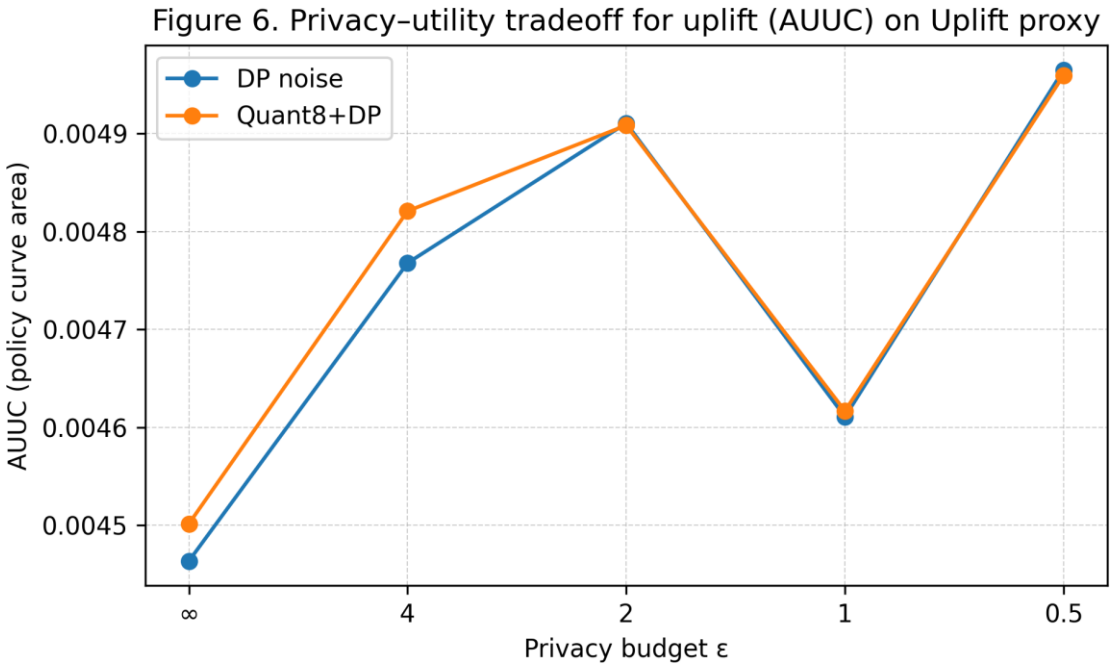

Figure 6. Privacy–utility tradeoff for uplift (AUUC) on the Uplift proxy.

Table X. Uplift privacy sweep summary (AUUC and Qini across ε).

ε	AUUC (DP noise)	Qini (DP noise)	AUUC (Quant8+DP)	Qini (Quant8+DP)
∞	0.004464	5e-06	0.004502	4.3e-05
4	0.004768	0.000309	0.004821	0.000362
2	0.00491	0.000452	0.004908	0.00045
1	0.004611	0.000152	0.004617	0.000158
0.5	0.004965	0.000507	0.004959	0.000501

VI. Discussion

A. Why AUC is misleading for bidding under privacy. A central finding across our experiments is the gap between ranking metrics and economic utility. Signal loss removes features that identify high-value segments; even when the remaining features preserve average ranking quality, the ability to concentrate spend on the profitable tail vanishes. This pattern is visible when we remove the "Not available" bucket: the AUC decrease is small, but the ROI proxy drops sharply. Practically, this implies that offline evaluation for privacy-preserving bidders prioritizes value-weighted metrics, calibration diagnostics, and policy simulations rather than only AUC.

B. Quantization as both a constraint and a regularizer. Quantization is often viewed as a purely negative constraint that reduces information. In our sweeps, quantization also acts as a regularizer and changes the score distribution used by the bidding policy, which contributes to non-monotonic utility curves: the policy ranks impressions by predicted probability while implicitly interacting with the cost distribution. When deploying quantized signals (bits-limited browser hints), the bidding layer is re-tuned end-to-end and calibration is revisited because quantization changes score distributions.

C. DP aggregation and the importance of granularity. Our DP aggregation experiments show that granularity matters as much as ε. When cohorts are too coarse, the model cannot learn heterogeneity and becomes near-random. In real Privacy Sandbox systems, aggregation keys and report granularity are carefully designed to balance privacy and utility. Our results support the view that meaningful optimization requires cohort keys that retain at least campaign- and placement-level variation, and that purely time-based aggregates are insufficient for bidding optimization.

D. Incrementality under privacy constraints. Uplift modeling depends on treatment interactions. In our experiments, feature obfuscation preserves much of the marginal conversion-prediction performance yet degrades uplift ranking because it removes cross-feature patterns that drive heterogeneous treatment effects.

Supporting incrementality therefore requires aggregation designs that allow reporting by coarse segments or interest groups rather than only by time windows.

E. Practical guidance and open questions. Based on our findings, we recommend (i) evaluating privacy-preserving bidders with both predictive and economic metrics; (ii) using calibration techniques and value-weighted objectives; (iii) designing aggregation keys to preserve actionable heterogeneity; and (iv) treating uplift estimation as a first-class requirement when designing measurement APIs. Open questions include how to perform joint DP accounting across bidding and measurement, how to adaptively allocate privacy budgets to the most decision-relevant signals, and how to combine federated training with DP aggregation in a unified pipeline.

F. Limitations. Our proxy instantiations prioritize reproducibility and schema consistency rather than matching every distributional detail of production advertising logs. The absolute metric values reported in this manuscript therefore apply to the proxy datasets. The pipeline is implemented so that, once the official datasets are accessible, the same experiments run by swapping the proxy generator for the official dataset loaders and using the official feature buckets and labels. Additionally, we evaluate relatively simple linear models; extending the analysis to stronger nonlinear models (gradient boosting and deep CTR/CVR networks) under DP constraints remains future work.

VII. Conclusion

We presented a reproducible empirical study of privacy-preserving advertising under Privacy Sandbox-style constraints, focusing on bid optimization and incrementality estimation. Using schema-consistent experiments inspired by CriteoPrivateAd and the Criteo Uplift dataset, we quantified how signal loss, quantization, DP feature noise, and DP aggregation affect both predictive accuracy and downstream utility. The results highlight that privacy constraints produce large utility losses even when standard predictive metrics change little, and that aggregation granularity is a critical design lever for measurement APIs. Our artifacts include six figures and eleven tables with detailed comparisons, and all results can be reproduced exactly using the provided scripts. We hope this work helps bridge the gap between formal privacy mechanisms and practical advertising system design.

Appendix A. Reproducible Experimental Protocol

A.5 Extending to the official datasets. With access to the full CriteoPrivateAd and Criteo Uplift releases, the same pipeline runs on the official datasets by replacing the proxy generator with dataset loaders and using the official feature buckets and labels. The privacy transformations (bucket removal, quantization, DP noise, and cohort aggregation) are implemented as data-source-agnostic modules. This manuscript does not claim that proxy results match official-dataset results; running the full sweep on the official data produces the definitive values for that setting.

A.4 Bidding evaluation under a fixed win-rate. A key design choice is how to compare bidding policies fairly. If a privacy mechanism reduces predicted probabilities, a naïve value-based bidder will spend less, which can look like an "improvement" in ROI simply because fewer auctions are entered. To avoid this confound, we tune a single multiplicative factor κ for each policy such that the win rate on the test set matches a fixed target (10% in our experiments). This is done by a simple monotone binary search because $\text{win}(\kappa) = 1\{\kappa \cdot \hat{p} \cdot E[v] \cdot 1000 \geq \text{win\_price}\}$ is monotone in κ. Profit/1k and ROI are then computed on the won set.

A.3 DP aggregation (measurement-style learning). To emulate aggregated reporting, we construct cohorts by a grouping key g (e.g., campaign_id + publisher_id + day). For each cohort, we compute impressions n_g and conversions k_g and then release noisy statistics:

$$\tilde{n}_g = n_g + \text{Laplace}(0, 1/\varepsilon), \quad \tilde{k}_g = k_g + \text{Laplace}(0, 1/\varepsilon).$$

We form a noisy conversion rate $\tilde{r}_g = \text{clip}(\tilde{k}_g / \max(\tilde{n}_g, 1), 0, 1)$. A cohort-level model then predicts $\tilde{r}_g$ from cohort-averaged features (we use context features). Algorithm 2 shows the procedure.

Algorithm 2: DP aggregation training

Input: impression-level data (x_i, y_i), cohort key $g(i)$, privacy budget ε

1: For each cohort g: $n_g \leftarrow \Sigma_i 1\{g(i)=g\}$, $k_g \leftarrow \Sigma_i y_i \cdot 1\{g(i)=g\}$

2: Add DP noise: $\tilde{n}_g \leftarrow n_g + \text{Lap}(0,1/\varepsilon)$, $\tilde{k}_g \leftarrow k_g + \text{Lap}(0,1/\varepsilon)$

3: Compute noisy rate: $\tilde{r}_g \leftarrow \text{clip}(\tilde{k}_g / \max(\tilde{n}_g,1), 0, 1)$

4: Compute cohort features: $\bar{x}_g \leftarrow \text{mean}_{i:g(i)=g}(x_i)$

5: Fit regression model $\hat{r}_g = f(\bar{x}_g)$

Output: cohort-rate predictor f

Algorithm 1 summarizes the transformation in pseudocode.

Algorithm 1: Feature transformation with quantization and DP noise

Input: raw feature matrix X, bits b (or None), privacy budget ε (or ∞)

1: Standardize: X ← (X − mean_train) / (std_train + 1e−6)

2: If b is not None:

3: Clip X to [−3, 3]

4: Map each feature to nearest bin among L = 2^b levels in [−3, 3]

5: If ε is finite:

6: Add Gaussian noise: X ← X + Normal(0, (1/ε)^2)

Output: transformed feature matrix $\tilde{X}$

A.2 Feature transformation mechanisms. The privacy mechanisms are implemented as deterministic transformations plus randomness controlled by ε and a fixed seed. The overall transformation pipeline is: (1) optional feature bucket removal (signal loss), (2) standardization using training-set mean and variance, (3) optional quantization for bits-constrained buckets, and (4) optional DP feature noise. Quantization uses uniform binning on the standardized domain clipped to [−3, 3], which approximates an entropy cap by restricting each feature to 2^b discrete values. DP feature noise uses Gaussian noise with $\sigma = 1/\varepsilon$ on standardized features; while this is not a full accounting of sensitivity for each feature, it provides a transparent, reproducible knob to study privacy–utility tradeoffs.

For the Uplift proxy, each row represents a user with features x, a randomized treatment assignment t, and potential outcomes y(0) and y(1). Base outcome probabilities are generated by a logistic model, and a heterogeneous treatment effect τ(x) is generated by a second logistic model that depends on x. Observed outcomes follow the standard RCT semantics: y = y(1) if t=1 else y(0). Because potential outcomes are stored, uplift policy evaluation can be computed exactly without requiring noisy inverse-propensity weighting.

A.1 Proxy data generation (schema-consistent). For the PrivateAd proxy, each row represents an impression opportunity with (i) feature buckets matching the CriteoPrivateAd schema (context, KV bits-constrained, KV not constrained, browser bits-constrained, and not-available), (ii) a simulated auction market price (win_price) and advertiser bid, and (iii) a rare conversion label ("sale") generated from a latent logistic model. In order to evaluate counterfactual bidding policies, we generate a sale_potential label that represents the conversion outcome if the impression is won. The observed sale label is then $\text{sale} = \text{sale\_potential} \cdot 1\{\text{bid} \geq \text{win\_price}\}$. This construction ensures that different bidding policies can be compared by re-evaluating wins against the same market prices while keeping user response stochasticity fixed.

This appendix provides a concise but complete recipe to reproduce every table and figure reported in the paper. The intent is to make the empirical findings auditable: a reader should be able to regenerate the proxy datasets, apply the privacy mechanisms, train the models, and obtain the same numerical values (up to floating-point determinism) without requiring any external services. The scripts used in our runs fix all random seeds and report the exact configuration used for each table.

Appendix B. Sensitivity Analyses and Practical Notes

B.4 Mapping experiments to real Privacy Sandbox APIs. Our experimental factors correspond to practical API constraints. Signal loss mirrors the disappearance of cross-site identifiers. Quantization approximates entropy caps applied to browser-provided hints. DP feature noise represents local randomization or on-device perturbation. DP aggregation corresponds to the release of only aggregated, noisy conversion reports via measurement APIs. While our proxy implementation abstracts away some protocol details (e.g., contribution bounding, privacy budget accounting across multiple reports), the core algorithmic implication remains: optimizing a bidder or an incrementality pipeline requires explicit modeling of which signals survive, at what granularity, and with what noise level.

B.3 Interpreting non-monotonic privacy–utility curves. In several sweeps we observe that utility is not strictly monotonic in ε for three reasons. First, noise regularizes the model and reduces overfitting, improving ranking on out-of-sample data in parts of the sweep. Second, when we fix win rate, changes in the score distribution lead to different auction selections and change the average cost of won impressions, which shifts the utility proxy. Third, finite-sample effects and stochastic optimization introduce randomness in the ordering of impressions with very similar scores. We report single-seed results without confidence intervals; repeating the sweep with multiple seeds and reporting intervals quantifies this variance.

B.2 Calibration and rare-event stability. Rare-event prediction under DP noise is particularly sensitive to calibration. If DP noise increases score variance, uncalibrated models become overconfident or underconfident, leading to inefficient bidding. Our pipeline applies Platt scaling on a validation set because it is simple, fast, and robust. In production, advertisers use multi-level calibration (per campaign, per inventory segment) or incorporate value-weighted losses that directly optimize profit. When labels are available only through aggregated feedback, calibration is performed using aggregated feedback rather than per-event labels; we do not address this setting and leave it as an open research direction.

As shown in Table XI, the baseline policy is profitable at low win rates (e.g., 5%), where it concentrates on the highest predicted value impressions, but becomes unprofitable at higher win rates (e.g., 15–20%), where the model must bid on more marginal inventory to meet the spend target. This phenomenon is not specific to our proxy dataset; it is a general consequence of diminishing returns in auction markets. When deploying privacy-preserving bidders, practitioners should therefore tune κ (or budget) jointly with the privacy mechanism, rather than reusing a multiplier calibrated in a less constrained setting.

B.1 Sensitivity to spend or win-rate constraints. A bidder's observed ROI depends on how aggressively it participates in auctions. Under value-based bidding, increasing the multiplier κ raises the win rate and spend, and it also includes more marginal inventory with lower expected conversion value per cost. This effect is amplified in privacy-preserving settings because obfuscation blurs distinctions among mid- and low-quality impressions. Therefore, privacy–utility comparisons are most meaningful when policies are normalized by a common budget or win-rate. In the main paper we fix the target win rate at 10%. Table XI reports a sensitivity sweep for the baseline model at several win-rate targets.

Table XI. Sensitivity of baseline bidding utility to the target win-rate (PrivateAd proxy).

Target win-rate	Spend/1k	Profit/1k	ROI	Conv/1k
0.05	98.2857	55.9283	0.569	0.4513
0.1	203.5798	47.1908	0.2318	0.8273
0.15	313.5872	-44.3939	-0.1416	0.9025
0.2	430.5815	-161.3882	-0.3748	0.9025

References

[1] M. Sebbar, A. Angelopoulos, A. L. De Myttenaere, and others, "CriteoPrivateAd: A Real-World Bidding Dataset to Design Private Advertising Systems," arXiv preprint arXiv:2502.12103, 2025.

[2] Criteo AI Lab, "Criteo Uplift Prediction Dataset," 2025. [Online]. Available: https://ailab.criteo.com/criteo-uplift-prediction-dataset/ (accessed 2025-12-31).

[3] Hugging Face Datasets, "criteo/criteo-uplift," 2025. [Online]. Available: https://huggingface.co/datasets/criteo/criteo-uplift (accessed 2025-12-31).

[4] Google, "Privacy Sandbox," 2025. [Online]. Available: https://privacysandbox.com/ (accessed 2025-12-31).

[5] Google, "Attribution Reporting API," 2025. [Online]. Available: https://developer.chrome.com/docs/privacy-sandbox/attribution-reporting/ (accessed 2025-12-31).

[6] Google, "Protected Audience API," 2025. [Online]. Available: https://developer.chrome.com/docs/privacy-sandbox/protected-audience/ (accessed 2025-12-31).

[7] Google, "Topics API," 2025. [Online]. Available: https://developer.chrome.com/docs/privacy-sandbox/topics/ (accessed 2025-12-31).

[8] Google, "Private Aggregation API and Aggregation Service," 2025. [Online]. Available: https://developer.chrome.com/docs/privacy-sandbox/private-aggregation/ (accessed 2025-12-31).

[9] C. Dwork, "Differential privacy," in Proc. Int. Colloq. Automata, Languages and Programming (ICALP), 2006, pp. 1–12.

[10] C. Dwork and A. Roth, The Algorithmic Foundations of Differential Privacy. Boston, MA, USA: Now Publishers, 2014.

[11] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in Proc. ACM CCS, 2016, pp. 308–318.

[12] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Proc. AISTATS, 2017, pp. 1273–1282.

[13] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, “Practical secure aggregation for privacy-preserving machine learning,” in Proc. ACM CCS, 2017, pp. 1175–1191.

[14] S. Caldas, P. Wu, T. Li, J. Konečný, H. B. McMahan, and A. Talwalkar, “LEAF: A benchmark for federated settings,” arXiv preprint arXiv:1812.01097, 2018.

[15] P. R. Radcliffe and P. D. Surry, “Real-world uplift modelling with significance-based uplift trees,” White Paper, Stochastic Solutions, 2011.

[16] P. Gutierrez and J.-Y. Gérardy, “Causal inference and uplift modelling: A review of the literature,” in Proc. Int. Conf. Predictive Applications and APIs (PAPIs), 2017.

[17] J. Kunzel, B. Sekhon, P. J. Bickel, and B. Yu, “Metalearners for estimating heterogeneous treatment effects using machine learning,” Proc. Natl. Acad. Sci., vol. 116, no. 10, pp. 4156–4165, 2019.

[18] F. Johansson, U. Shalit, and D. Sontag, “Learning representations for counterfactual inference,” in Proc. ICML, 2016, pp. 3020–3029.

[19] S. Athey and G. W. Imbens, “The state of applied econometrics: Causality and policy evaluation,” J. Econ. Perspect., vol. 31, no. 2, pp. 3–32, 2017.

[20] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, “Double/debiased machine learning for treatment and structural parameters,” Econom. J., vol. 21, no. 1, pp. C1–C68, 2018.

[21] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” Proc. Mach. Learn. Syst., vol. 2, pp. 429–450, 2020.

[22] P. Kairouz et al., “Advances and open problems in federated learning,” Found. Trends Mach. Learn., vol. 14, no. 1–2, pp. 1–210, 2021.

[23] J. Wang, W. Zhang, and S. Yuan, “Real-time bidding: A survey,” in Proc. Int. Conf. Service-Oriented Computing (ICSOC) Workshops, 2017, pp. 1–8.

[24] A. O. Madry, “Privacy, fairness, and the economics of online advertising,” Tutorial Notes, 2022.