Contract Cards for Auditable Private Conformal Prediction

Patrick Indri [1] **Tamara Drucks** [1 2] **Georgios Spathoulas** [3 4]

Abstract

We propose contract cards for auditable private conformal prediction. Private conformal methods provide coverage and privacy guarantees, but deployment audit also requires explicit records of which requirements were imposed, which configuration was selected, and what evidence supports the decision. A contract card pairs user-specified coverage and privacy requirements with either an infeasibility report or a certified private conformal configuration, together with diagnostics that describe operational cost. We instantiate the idea with a transparent differentially private adaptive prediction set backend. This instantiation supplies the formal quantities recorded in the card: a finite-sample coverage lower bound, a calibration privacy guarantee, and a certificate-width diagnostic that tracks privacy-induced threshold inflation. Experiments on image-classification data show that the recorded fields satisfy the formal checks and vary predictably with calibration privacy and sample size, demonstrating that contract cards provide useful evidence for lightweight audit.

1. Introduction

Trustworthy model *deployment* requires evidence that can be inspected after a model is released. Aggregate accuracy is rarely enough: deployment decisions may depend on whether uncertainty estimates have a valid coverage guarantee, whether sensitive records were protected during training and calibration, and whether later auditors can reconstruct the basis for the decision. Deployed machine learning (ML) models are also increasingly expected to be transparent, private, and auditable (European Parliament and Council, 2024; 2016; NIST, 2023). Conformal prediction (Shafer and Vovk, 2008) provides finite-sample uncertainty guarantees, while differential privacy (DP; Dwork, 2006) gives a formal guarantee for bounding information leakage. Private conformal prediction is therefore a natural building block for trustworthy ML systems. Guarantees alone, however, do not make a deployed predictor easy to audit. Existing private conformal methods (Angelopoulos et al., 2022; Romanus and Molinari, 2025; Wu et al., 2026) are usually presented as calibration algorithms: they prove privacy and coverage for a fixed private calibration procedure and report empirical efficiency for selected hyperparameters. Deployment review asks a different question: *given explicit coverage and privacy requirements, which configuration was selected, why was it admissible, and what evidence was retained so the decision can be checked later?* This question becomes especially important when requirements change. Auditors and practitioners may want to know whether tighter coverage or privacy targets remain feasible, or whether additional calibration data may improve the uncertainty estimates. Thus, useful audit evidence should make changes in feasibility and efficiency visible rather than implicit.

To this end, we propose the use of *contract cards*: compact records attached to private conformal predictors. A card contains the requested coverage–privacy contract, the selected configuration, the formal quantities needed to recompute feasibility, and diagnostics such as empirical coverage and prediction-set size. Unlike broad documentation artifacts such as model cards, datasheets, or factsheets (Mitchell et al., 2019; Gebru et al., 2021; Arnold et al., 2019), contract cards have a narrower scope: they make the coverage and privacy claims of a private conformal predictor explicit and checkable. We instantiate this idea with a transparent differentially private adaptive prediction set backend (APS; Romano et al., 2020). The backend is deliberately simple but exposes the audit-relevant quantities: coverage and privacy guarantees, a certificate-width bound quantifying DP-induced threshold inflation, and predictable variation with calibration privacy and sample size. Together, these ingredients give a concrete implementation of contract cards for socially consequential deployments, including public-sector and other high-stakes settings, separating formal feasibility checks from empirical diagnostics and making privacy–coverage–efficiency trade-offs visible and auditable.

[1]Research Unit Machine Learning, TU Wien, Vienna, Austria [2]Faculty of Computer Science, University of Vienna, Vienna, Austria [3]Department of Information Security and Communication Technology, NTNU, Gjøvik, Norway [4]Department of Computer Science and Biomedical Informatics, University of Thessaly, Thessaly, Greece. Correspondence to: Patrick Indri <patrick.indri@tuwien.ac.at>.

Trustworthy AI for Good (AI4GOOD) Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

2. Related work

Private conformal prediction combines split conformal calibration with private quantile or threshold selection, using mechanisms such as conservative private quantiles (Angelopoulos et al., 2022), randomized quantile binary search (Romanus and Molinari, 2025), and statistically efficient non-splitting calibration (Wu et al., 2026), building on private quantile estimation more broadly (Smith, 2011; Gillenwater et al., 2021; Kaplan et al., 2022; Durfee, 2023). These methods primarily optimize the privacy–utility trade-off of the calibration rule, whereas our focus is on how a deployment-facing system should select, reject, and document a private conformal configuration against explicit coverage and privacy requirements. Our contract view is also related to distribution-free risk control, learn-then-test, and conformal risk control, which calibrate predictive systems to satisfy statistical requirements (Bates et al., 2021; Angelopoulos et al., 2025; 2024). We add differential privacy and focus on the retained evidence: unlike broad documentation artifacts such as model cards, datasheets, or factsheets (Mitchell et al., 2019; Gebru et al., 2021; Arnold et al., 2019), contract cards are narrower records that contain the formal quantities needed to check a stated coverage–privacy claim and diagnostics for interpreting its operational cost.

3. Preliminaries

We use upper-case letters for random variables and lower-case letters for their realizations when the distinction is important. We use the same notation for functions evaluated on random variables or on realized values when the meaning is clear from context. All missing proofs are reported in Section A. We consider a multiclass classification setting with input–label pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the space of inputs with labels $\mathcal{Y} = \{1, \dots, K\}$. A set-valued predictor is a function $C : \mathcal{X} \to 2^{\mathcal{Y}}$, where $2^{\mathcal{Y}}$ is the collection of subsets of $\mathcal{Y}$. The *coverage* of a set-valued predictor on a test point x is $\mathbb{P}\{Y \in C(X)\}$. Conformal prediction (Shafer and Vovk, 2008) constructs such predictors from a held-out calibration set $(X_i, Y_i)_{i=1}^m$ and relies on the assumption of *exchangeability*: a sequence of random variables is exchangeable if its joint distribution is invariant under permutation. Given a score function $R : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and a threshold $\tau \in \mathbb{R}$, the prediction set is defined as $C_\tau(x) = \{y \in \mathcal{Y} : R(x, y) \le \tau\}$. For calibration scores $R_i = R(X_i, Y_i)$, let the *nominal coverage* be γ, and $k = \lceil (m+1)\gamma \rceil$, the split conformal threshold is the kth quantile of $R_1, \dots, R_m$ when $k \le m$. Under exchangeability of the calibration data and test point, the resulting predictor has marginal coverage at least $k/(m+1) \ge \gamma$. In our instantiation, R is the adaptive prediction set score (APS; Romano et al., 2020). Let $\hat{p}(\cdot \mid x)$ be the classifier's predictive distribution, let $p_{(1)}(x) \ge \cdots \ge p_{(K)}(x)$ denote the sorted class probabilities, and let $J(x, y)$ be the rank of label y in this ordering. The APS score is $R(x, y) = \sum_{r=1}^{J(x,y)} p_{(r)}(x)$. The sets $C_\tau(x)$ are monotone in τ.

Differential privacy (DP; Dwork, 2006; Dwork and Roth, 2014) formalizes protection against the effect of changing one data point. A *database* is a finite collection of points, and two databases D, D' are *adjacent* under add/remove adjacency if one can be obtained from the other by adding or removing a single point.

Definition 3.1 ((ϵ, δ)-DP, Dwork and Roth 2014)**.** Let $\epsilon \ge 0$ and $\delta \ge 0$. A randomized mechanism $\mathcal{M}$ with range $\mathrm{Range}(\mathcal{M})$ satisfies (ϵ, δ) differential privacy, denoted (ϵ, δ)-DP, if, for all adjacent databases D, D' and all measurable sets $S \subseteq \mathrm{Range}(\mathcal{M})$,

$$\Pr[\mathcal{M}(D) \in S] \le e^{\epsilon} \Pr[\mathcal{M}(D') \in S] + \delta, \qquad (1)$$

where probabilities are taken over the randomness of $\mathcal{M}$. When $\delta = 0$, we say that $\mathcal{M}$ satisfies *pure* ϵ-DP.

We refer to ϵ as the privacy budget and to δ as the failure parameter. In our setting, $\epsilon_{\mathrm{train}}$ denotes the training privacy budget and ϵ_{cal} denotes the calibration privacy budget.

We now introduce the terminology used in the paper. A *contract* is a tuple $\kappa = (\gamma_{\mathrm{target}}, \bar{\epsilon}_{\mathrm{train}}, \bar{\epsilon}_{\mathrm{cal}}, \beta)$, where γ_{target} is the requested coverage level, $\bar{\epsilon}_{\mathrm{train}}$ and $\bar{\epsilon}_{\mathrm{cal}}$ are maximum allowed privacy budgets, and β is the failure probability used in the coverage certificate. A *configuration* is a tuple $c = (\gamma, \epsilon_{\mathrm{train}}, \epsilon_{\mathrm{cal}}, \xi)$, where ξ contains auxiliary choices such as calibration size, grid resolution, score function, and model family, as well as a fixed privacy parameter δ. A configuration c induces a set-valued predictor $C(\cdot; c)$. A *certificate* is formal evidence used to check whether c satisfies κ, including a theorem-certified lower bound $L(c)$ such that $\mathbb{P}\{Y \in C(X; c)\} \ge L(c)$. A *diagnostic* is an empirical quantity, such as observed coverage or average prediction-set size, used to describe the behavior of the model. A *contract card* records the contract, the selected configuration or infeasibility report, the certificate quantities needed to recompute formal feasibility, and diagnostics to describe the privacy–coverage–efficiency trade-off.

4. Contract Cards as Audit Artifacts

The role of the contract card is to make the contract, the certificate, and the diagnostics visible in one artifact. Rather than reporting only, e.g., empirical coverage estimates, the contract card records the quantities needed to recompute formal feasibility (Equation (3)) and the diagnostics needed to interpret what changes when requirements are varied. Given a contract κ and configuration c, suppose that the calibration procedure provides a theorem-certified lower bound $L(c)$ satisfying

$$\mathbb{P}\{Y \in C(X; c)\} \ge L(c). \qquad (2)$$

We say that c is *formally feasible* for κ when

$$\mathrm{Feas}(c;\kappa) \iff [L(c) \geq \gamma_{\mathrm{target}}] \wedge [\epsilon_{\mathrm{train}} \leq \bar{\epsilon}_{\mathrm{train}}] \wedge [\epsilon_{\mathrm{cal}} \leq \bar{\epsilon}_{\mathrm{cal}}]\,. \tag{3}$$

When training is $(\epsilon_{\mathrm{train}}, \delta)$-DP and calibration is pure ϵ_{cal}-DP, standard composition gives an overall $(\epsilon_{\mathrm{train}} + \epsilon_{\mathrm{cal}}, \delta)$ privacy guarantee (Dwork and Roth, 2014). The card keeps the two budgets separate because they answer different audit questions: how much privacy was spent fitting the predictor, and how much was spent calibrating its uncertainty sets. A contract card has two layers. The *formal layer* records the contract, configuration, lower bound, and feasibility flags needed to recompute Equation (3); see Observation 4.1. The *diagnostic layer* records empirical summaries such as observed coverage and average prediction-set size that help interpret the selected model but do not replace the formal guarantee. Table 1 gives a compact schema.

Observation 4.1 (Feasibility Check)**.** *Fix a contract κ and a configuration c. Assume a card records correct values for κ, c, $L(c)$, and the privacy coordinates of c. Then any verifier that recomputes Equation (3) from the card accepts if and only if the selected configuration satisfies the stated formal coverage and privacy guarantees.*

Let $\mathcal{G}$ denote the finite configuration grid searched by the deployment procedure. Each element $c \in \mathcal{G}$ is a candidate configuration of the form $c = (\gamma, \epsilon_{\mathrm{train}}, \epsilon_{\mathrm{cal}}, \xi)$, where the coordinates range over user-specified finite sets, for example nominal coverage levels, privacy budgets, calibration sizes, grid resolutions, and model families. The grid $\mathcal{G}$ is therefore not a statistical object; it is the design space over which the deployment procedure is allowed to search. For each candidate $c \in \mathcal{G}$, the procedure evaluates $\mathrm{Feas}(c;\kappa)$ using the certificate quantities available before empirical testing. If at least one configuration is formally feasible, the card records the selected configuration, the number of configurations checked, and the subset of configurations that are formally feasible. If no configuration is formally feasible, the card reports formal infeasibility and the failed margin, for example $\max_{c\in\mathcal{G}} L(c) - \gamma_{\mathrm{target}}$. Diagnostics are reported for configurations that pass the formal check; observed coverage or set size can guide selection among formally feasible configurations but not make a formally infeasible configuration feasible.

5. A Transparent Private Conformal Backend

We now describe the backend used to instantiate the card. The backend is intentionally transparent rather than optimized for the best possible privacy–utility trade-off: it exposes how the private threshold and its certificate change with calibration privacy, calibration size, and nominal coverage. Let $(X_1, Y_1), \ldots, (X_m, Y_m)$ be the calibration data and $R_i = R(X_i, Y_i)$ the APS calibration scores defined above. For nominal coverage γ, let $k = \lceil (m+1)\gamma \rceil$. To calibrate privately, we replace the exact quantile with noisy cumulative counts on a public grid. Fix grid points $t_b = b/B$ for $b \in \{1, \ldots, B\}$, $B \in \mathbb{N}$, and define cumulative counts $N_b = \sum_{i=1}^{m} \mathbf{1}\{R_i \leq t_b\}$. The count vector has ℓ_1 sensitivity at most B under add/remove adjacency. We release noisy counts $\widetilde{N}_b = N_b + Z_b$, with $Z_b \sim \mathrm{Lap}(B/\epsilon_{\mathrm{cal}})$ iid, and use the conservative offset $\lambda = B \log(B/\beta)/\epsilon_{\mathrm{cal}}$. The private threshold is $\hat{\tau} = t_{\hat{b}}$, where $\hat{b} = \min\{b : \widetilde{N}_b \geq k + \lambda\}$ with the convention $\hat{b} = B$ if the set is empty. We now prove that a backend constructed as described has provable conformal coverage and DP guarantees.

Theorem 5.1 (DP-APS calibration)**.** *Assume the calibration data and test point are exchangeable and the classifier is fixed before calibration. The noisy-count mechanism that releases $\hat{\tau}$ is pure ϵ_{cal}-DP under add/remove adjacency. Moreover, with probability at least $1-\beta$, the selected threshold is at least as conservative as the non-private conformal grid threshold, and the resulting APS predictor satisfies $\mathbb{P}\{Y_{m+1} \in C_{\hat{\tau}}(X_{m+1})\} \geq \gamma - \beta$ for an unseen test point (X_{m+1}, Y_{m+1}).*

Theorem 5.1 shows that the backend records the certified lower bound $L(c) = \gamma - \beta$. The proof, reported in Section A, also identifies the privacy–efficiency trade-off. Decreasing ϵ_{cal} gives stronger privacy, but it increases the Laplace scale $B/\epsilon_{\mathrm{cal}}$ and the conservative offset $\lambda = (B/\epsilon_{\mathrm{cal}}) \log(B/\beta)$. A larger λ raises the noisy-count threshold $k + \lambda$, so the selected threshold is typically larger and the resulting prediction sets contain more labels. A contract card makes this mechanism visible by recording both the formal lower bound and diagnostics affected by privacy noise. The most direct such diagnostic is the *certificate width* $W(c)$, which bounds the possible inflation of the private threshold relative to the non-private grid threshold.

Proposition 5.2 (Threshold-inflation certificate)**.** *Define the empirical grid-quantile index map $q(r) = \min\{b : N_b \geq r\}$, with $q(r) = B$ if the set is empty. With probability at least $1-\beta$, the private threshold $\hat{\tau}$ satisfies $t_{q(k)} \leq \hat{\tau} \leq t_{q(\lceil k+2\lambda \rceil)}$. Consequently, the* certificate width $W(c) = t_{q(\lceil k+2\lambda \rceil)} - t_{q(k)}$ *upper-bounds the inflation of the private threshold relative to the non-private grid threshold.*

For monitoring, larger $W(c)$ can indicate that stronger privacy, smaller calibration samples, or sparse calibration scores near the target quantile are making the private uncertainty sets less efficient. Since $W(c)$ is data-dependent, public release requires either a trusted audit setting or additional privacy accounting; in our experiments, it is reported as a diagnostic of the card.

Remark 5.3 (Monotone certificate behavior). For fixed γ, β, B, and empirical counts $(N_b)_{b=1}^{B}$, the conservative offset $\lambda = (B/\epsilon_{\mathrm{cal}}) \log(B/\beta)$ is decreasing in ϵ_{cal}. Thus, increasing the calibration privacy budget weakens privacy but

Table 1. Contract-card fields. Formal fields support recomputation of the coverage–privacy predicate; diagnostic fields support comparison and monitoring.

Field group	Example field	Audit role
Contract	$\gamma_{\text{target}}, \bar{\epsilon}_{\text{train}}, \bar{\epsilon}_{\text{cal}}, \beta$	Deployment requirement
Configuration	$\gamma, \epsilon_{\text{train}}, \epsilon_{\text{cal}}, \xi$	Selected calibration choices
Formal certificate	$L(c)$ and feasibility flags	Recompute checks
Certificate diagnostics	width $W(c)$, observed inflation	Quantify private conservativeness
Empirical diagnostics	coverage, set size, accuracy	Interpret operational behavior
Selection summary	feasible count, checked count, decision	Explain selection or infeasibility

can only decrease the upper index $q(\lceil k + 2\lambda \rceil)$ in Proposition 5.2. The certificate width $W(c)$ is therefore monotone nonincreasing in ϵ_{cal} for fixed calibration scores and grid.

This monotone certificate behavior makes the card more informative: it shows not only which configuration was selected, but also how nearby privacy choices would change the efficiency certificate.

6. Auditing and Monitoring

Contract cards support three tasks that arise in deployment.

Formal audit. An auditor recomputes Equation (3) from the card. This verifies that the selected configuration meets the recorded coverage lower bound and privacy–budget constraints. If the card reports infeasibility, the same fields identify which clause failed and by what margin.

Failure-mode analysis. The formal and diagnostic fields indicate why a configuration may fail certification or be operationally poor despite being formally valid. If $L(c) < \gamma_{\text{target}}$, the requested coverage cannot be certified on the searched grid. If ϵ_{cal} is too small, the threshold-inflation certificate can widen and prediction sets can become too large. If the base classifier is weak, coverage can remain formally valid while the average set size becomes impractical.

Sensitivity analysis. Cards can be regenerated under different coverage targets, calibration privacy budgets, or calibration-set sizes. Comparing them shows how formal feasibility and operational cost change with the requirements: for example, whether smaller ϵ_{cal} widens the threshold-inflation certificate or increases average set size. This gives auditors a compact record of the predictor's coverage–privacy evidence and diagnostics, which can be eventually combined with other evaluations for fairness, robustness, or other desiderata.

7. Experiments

We evaluate whether the card fields behave predictably and expose the intended privacy–coverage–efficiency trade-offs. The goal is not to produce state-of-the-art private prediction sets, but to test whether the audit record contains useful formal and empirical evidence about how configurations respond to changed requirements. We use CIFAR-10 (Krizhevsky, 2009) as a proof-of-concept benchmark for the card fields. We rely on a lightweight convolutional classifier trained by DP-SGD (Abadi et al., 2016). Training uses $\epsilon_{\text{train}} = 4.0$, $\delta = 10^{-5}$, four epochs, temperature 2.0, and label smoothing 0.5. We vary $\epsilon_{\text{cal}} \in \{2, 4, 8\}$, nominal coverage $\gamma \in \{0.55, 0.65, 0.75, 0.85\}$, and calibration-set size in $\{1000, 2000, 4000\}$, averaging over 3 random seeds. Each run records empirical coverage, average prediction-set size, private threshold, observed threshold inflation, and certificate width.

Figure 1 shows the main contract card diagnostics. The certificate width tightens as ϵ_{cal} increases and as calibration size increases. The middle panel shows the corresponding efficiency effect: stronger privacy generally produces larger prediction sets. The right panel checks the certificate interpretation directly: observed threshold inflation remains below the certificate width. These plots are useful as audit evidence because they distinguish three quantities that are often conflated: the formal coverage lower bound, the operational cost of privacy in set size, and the realized conservativeness of the private threshold. The main takeaway is that the contract card fields vary in the expected direction: changing privacy and sample size changes the certificate and efficiency diagnostics visibly, while the formal coverage predicate remains separately checkable.

Table 2. Summary of certificate-aware search on CIFAR-10. γ_{target} is the requested coverage target, N is the number of grid configurations, N_{run} is the number of seed-level runs, N_{form} is the number of formally feasible configurations, N_{sel} is the number of formally feasible configurations selected for empirical evaluation, and N_{eval} is the number of seed-level empirical evaluations performed on those selected configurations.

γ_{target}	N	N_{run}	N_{form}	N_{sel}	N_{eval}	Red. grid	Red. form
0.6	36	108	27	3	9	91.7%	88.9%
0.7	36	108	18	3	9	91.7%	83.3%
0.8	36	108	9	3	9	91.7%	66.7%

Table 2 shows which configurations pass the formal feasibility check before empirical evaluation. As the requested coverage target tightens, fewer configurations are formally admissible: 27 for $\gamma_{\text{target}} = 0.6$, 18 for $\gamma_{\text{target}} = 0.7$, and 9

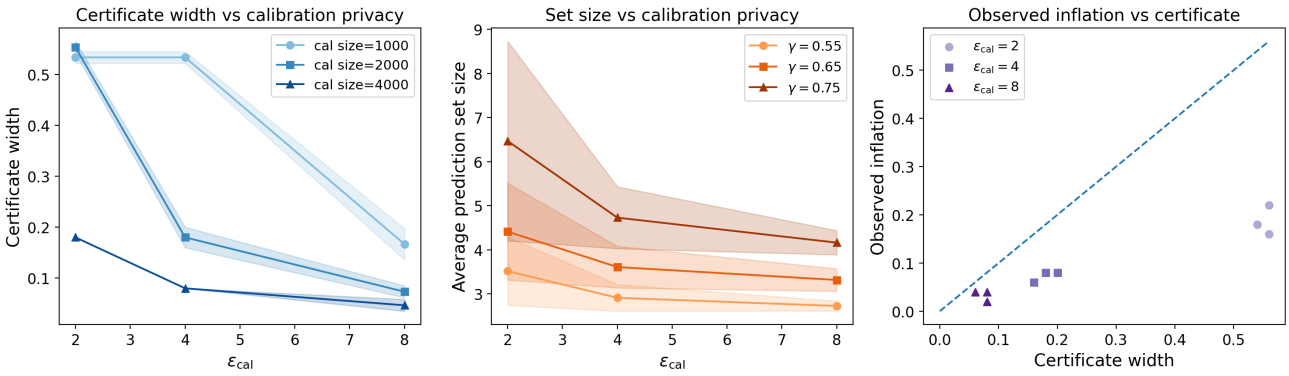

Figure 1. Main contract-card diagnostics on CIFAR-10. Left: certificate width decreases with larger calibration privacy budget and larger calibration sets. Middle: average prediction-set size decreases as ϵ_{cal} increases. Right: observed threshold inflation remains below the certificate width. Shaded regions show one standard deviation over random seeds.

for $\gamma_{\text{target}} = 0.8$. For each contract, we then evaluate only a small set of representative feasible configurations. For fixed calibration scores and grid, smaller ϵ_{cal} gives stronger privacy but cannot decrease the certificate width. Thus, once the card already contains a feasible configuration for a given privacy budget, configurations with the same contract and larger certificate width are less useful to evaluate empirically: they are expected to have at least as much privacy-induced conservativeness. This leaves 3 configurations per contract, or 9 seed-level evaluations instead of all 108 runs. The card records this information: which configurations were formally admissible, which were evaluated empirically, and why the final configuration was chosen.

> **Example contract card, abridged.** Request: $\gamma_{\text{target}} = 0.70$, $\bar{\epsilon}_{\text{train}} = 4$, $\bar{\epsilon}_{\text{cal}} = 8$, $\beta = 10^{-3}$. Decision: `FEASIBLE`. Selected configuration c: $\gamma = 0.75$, $\epsilon_{\text{train}} = 4$, $\epsilon_{\text{cal}} = 8$, $\delta = 10^{-5}$, calibration-set size $= 4000$. Formal evidence: $L(c) = 0.749 \geq 0.70$, train and calibration privacy checks pass. Certificate: $W(c) = 0.04$, observed threshold inflation $= 0.02$.

A formally infeasible card would instead report the best attempted configuration and the failed margin, for example a negative coverage margin if no searched configuration satisfies $L(c) \geq \gamma_{\text{target}}$.

Overall, the experiments show that contract cards provide a compact and interpretable record of both formal feasibility and the operational cost of satisfying a coverage–privacy contract. By recording certificate width, threshold inflation, and prediction-set size alongside the selected configuration, the card makes the result easier to inspect and compare.

8. Discussion and limitations

Contract cards make a narrow claim easier to inspect: under the stated assumptions and recorded configuration, a private conformal predictor satisfies a recomputable coverage–privacy predicate, and its diagnostics summarize the operational cost of that guarantee. While contract cards help model audit and monitoring, it should be highlighted that they do not constitute a complete trustworthiness certificate. A valid card does not establish fairness, robustness, calibration under arbitrary distribution shift, or the social appropriateness of a downstream use case. We leave the extension of a formal treatment of these desiderata for trustworthy ML as future work.

The current instantiation has three main limitations. First, DP-APS was chosen for transparency rather than efficiency. Stronger private quantile mechanisms could be substituted if they expose comparable contract-facing quantities, but not every private conformal method exposes certificate quantities whose behavior under changed requirements is easy to summarize. Second, the certificate width is data-dependent, so releasing it publicly may require a trusted audit setting or additional privacy accounting. Third, our experiments use image-classification benchmarks and a lightweight DP-trained classifier; they should be read as a proof of concept for the audit interface, not as deployment evidence for a high-stakes domain. The broader point is that trustworthy ML systems should retain structured evidence for formal uncertainty and privacy claims. Contract cards provide one compact way to do this for private conformal prediction.

Impact statement

This work aims to improve auditability for private conformal prediction. Its potential benefit is to make coverage and privacy claims easier to inspect, reproduce, and monitor after deployment. Its main risk is overinterpretation: a valid contract card does not imply that a deployed system is fair, robust, or socially appropriate. Contract cards should therefore be used as one component of a broader model evaluation and governance process.

Acknowledgments

This work was carried out under the "Provable Policy-Driven Privacy Compliance" action, which has received funding from the European Union, via the oc4-2025-TES-02 call issued and implemented by the ENFIELD project, under the grant agreement No 101120657. This work was also funded in part by the Vienna Science and Technology Fund (WWTF), project StruDL (ICT22-059); by the Austrian Science Fund (FWF), project NanOX-ML (6728).

References

European Parliament and Council. Artificial Intelligence Act, 2024. Regulation (EU) 2024/1689.

European Parliament and Council. General Data Protection Regulation, 2016. Regulation (EU) 2016/679.

NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical report, U.S. Department of Commerce, 2023.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer, 2006.

Anastasios Nikolas Angelopoulos, Stephen Bates, Tijana Zrnic, and Michael I Jordan. Private prediction sets. *Harvard Data Science Review*, 4(2), 2022.

Ogonnaya M Romanus and Roberto Molinari. Differentially private conformal prediction via quantile binary search. *arXiv preprint arXiv:2507.12497*, 2025.

Jiamei Wu, Ce Zhang, Zhipeng Cai, Jingsen Kong, Bei Jiang, Linglong Kong, and Lingchen Kong. Differentially private conformal prediction. *arXiv preprint arXiv:2604.14621*, 2026.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6–1, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 813–822, 2011.

Jennifer Gillenwater, Matthew Joseph, and Alex Kulesza. Differentially private quantiles. In *International Conference on Machine Learning*, pages 3713–3722. PMLR, 2021.

Haim Kaplan, Shachar Schnapp, and Uri Stemmer. Differentially private approximate quantiles. In *International Conference on Machine Learning*, pages 10751–10761. PMLR, 2022.

David Durfee. Unbounded differentially private quantile and maximum estimation. *Advances in Neural Information Processing Systems*, 36:77691–77712, 2023.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *ICLR*, 2024.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends in theoretical computer science*, 9(3-4):211–487, 2014.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/cifar.html.

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

A. Missing Proofs

We report here the missing proofs.

Observation 4.1 (Feasibility Check)**.** *Fix a contract κ and a configuration c. Assume a card records correct values for κ, c, $L(c)$, and the privacy coordinates of c. Then any verifier that recomputes Equation (3) from the card accepts if and only if the selected configuration satisfies the stated formal coverage and privacy guarantees.*

Proof. This conjunction is exactly the definition of formal feasibility, so verifier acceptance is equivalent to formal satisfaction of the contract by c. □

Theorem 5.1 (DP-APS calibration)**.** *Assume the calibration data and test point are exchangeable and the classifier is fixed before calibration. The noisy-count mechanism that releases $\hat{\tau}$ is pure ϵ_{cal}-DP under add/remove adjacency. Moreover, with probability at least $1-\beta$, the selected threshold is at least as conservative as the non-private conformal grid threshold, and the resulting APS predictor satisfies $\mathbb{P}\{Y_{m+1} \in C_{\hat{\tau}}(X_{m+1})\} \geq \gamma - \beta$ for an unseen test point (X_{m+1}, Y_{m+1}).*

Proof. For point (1), adding or removing one calibration point changes each cumulative count N_b by at most 1, so the vector $(N_1, \ldots, N_B)$ has ℓ_1 sensitivity at most B. Hence, by the vector Laplace mechanism, releasing

$$(\tilde{N}_1, \ldots, \tilde{N}_B) = (N_1, \ldots, N_B) + (Z_1, \ldots, Z_B),$$
$$Z_b \overset{\text{iid}}{\sim} \text{Lap}\left(\frac{B}{\epsilon_{\text{cal}}}\right),$$

is ϵ_{cal}-DP. Since $\hat{\tau}$ is a deterministic function of the noisy vector, it is also ϵ_{cal}-DP.

For point (2), define the good event $\mathcal{E} = \{|Z_b| \leq \lambda \text{ for all } b \in \{1, \ldots, B\}\}$, that is, the event that all additive noises are small. For Laplace noise with scale B/ϵ_{cal}, it holds that

$$\mathbb{P}(|Z_b| > \lambda) = \exp\left(-\frac{\epsilon_{\text{cal}}\lambda}{B}\right) = \frac{\beta}{B}.$$

Then, by union bound, the probability of the complement event $\mathcal{E}^c = \{\exists b : |Z_b| > \lambda\}$ can be bound as

$$\mathbb{P}(\mathcal{E}^c) \leq \sum_{b=1}^{B} \mathbb{P}(|Z_b| > \lambda) \leq B \cdot \frac{\beta}{B} = \beta.$$

Hence, $\mathbb{P}(\mathcal{E}) \geq 1 - \beta$. Assume now that $\mathcal{E}$ holds. If the set $\{b : \tilde{N}_b \geq k + \lambda\}$ is nonempty, then by definition of $\hat{b}$, it holds that $\tilde{N}_{\hat{b}} \geq k + \lambda$. Since $|Z_{\hat{b}}| \leq \lambda$ on $\mathcal{E}$, then

$$N_{\hat{b}} = \tilde{N}_{\hat{b}} - Z_{\hat{b}} \geq \tilde{N}_{\hat{b}} - |Z_{\hat{b}}| \geq (k + \lambda) - \lambda = k.$$

If the set is empty, then $\hat{b} = B$, so $\hat{\tau} = t_B = 1$, and therefore $N_B = m \geq k$. Thus in all cases, on $\mathcal{E}$ we have $N_{\hat{b}} \geq k$.

For point (3), first note the following deterministic implication: for every (x, y) and every threshold τ, it holds that $R(x, y) \leq \tau \implies y \in C_\tau(x)$. Indeed, this is immediate from the definition $C_\tau(x) = \{y : R(x, y) \leq \tau\}$. Let $A = \{Y_{m+1} \in C_{\hat{\tau}}(X_{m+1})\}$. By part (2), on the event $\mathcal{E}$ we have $N_{\hat{b}} \geq k$. Hence, at least k calibration scores are at most $\hat{\tau}$, so $\hat{\tau} \geq R_{(k)}$. Therefore,

$$\mathcal{E} \cap \{R_{m+1} \leq R_{(k)}\} \subseteq A.$$

It follows that

$$\mathbb{P}(A) \geq \mathbb{P}(\mathcal{E} \cap \{R_{m+1} \leq R_{(k)}\}).$$

Using the bound

$$\mathbb{P}(B \cap C) \geq \mathbb{P}(B) - \mathbb{P}(C^c),$$

we obtain

$$\mathbb{P}(A) \geq \mathbb{P}(R_{m+1} \leq R_{(k)}) - \mathbb{P}(\mathcal{E}^c).$$

By exchangeability, the multiset of scores $(R_1, \ldots, R_m, R_{m+1})$ is invariant under permutations. Let U denote the rank of R_{m+1} among these $m + 1$ values, with ties broken uniformly at random. Then U is uniformly distributed on $\{1, \ldots, m + 1\}$, and the event $\{U \leq k\}$ implies $\{R_{m+1} \leq R_{(k)}\}$. Therefore,

$$\mathbb{P}(R_{m+1} \leq R_{(k)}) \geq \mathbb{P}(U \leq k) \geq \frac{k}{m+1} \geq \gamma.$$

Moreover, by part (2), $\mathbb{P}(\mathcal{E}^c) \leq \beta$. Hence,

$$\mathbb{P}\left(Y_{m+1} \in C_{\hat{\tau}}(X_{m+1})\right) = \mathbb{P}(A) \geq \gamma - \beta.$$

This concludes the proof. □

Proposition 5.2 (Threshold-inflation certificate)**.** *Define the empirical grid-quantile index map $q(r) = \min\{b : N_b \geq r\}$, with $q(r) = B$ if the set is empty. With probability at least $1 - \beta$, the private threshold $\hat{\tau}$ satisfies $t_{q(k)} \leq \hat{\tau} \leq t_{q(\lceil k+2\lambda \rceil)}$. Consequently, the* certificate width *$W(c) = t_{q(\lceil k+2\lambda \rceil)} - t_{q(k)}$ upper-bounds the inflation of the private threshold relative to the non-private grid threshold.*

Proof. Consider the same good event as in the proof of Theorem 5.1, namely, the event $\mathcal{E} = \{|Z_b| \leq \lambda \text{ for all } b \in \{1, \ldots, B\}\}$. As in the proof of Theorem 5.1, $\mathbb{P}(\mathcal{E}) \geq 1-\beta$. Let us assume $\mathcal{E}$ happens.

As a first step, we intend to show that $q(k) \leq \hat{b}$. By definition, either $\hat{b} = B$ or else $\tilde{N}_{\hat{b}} \geq k + \lambda$. In the latter case, for the event $\mathcal{E}$, $|Z_{\hat{b}}| \leq \lambda$ and thus $N_{\hat{b}} = \tilde{N}_{\hat{b}} - Z_{\hat{b}} \geq k+\lambda-\lambda = k$. In the former case, $N_{\hat{b}} = N_B = m \geq k$. Thus, on $\mathcal{E}$, it always holds that $N_{\hat{b}} \geq k$ and therefore $q(k) \leq \hat{b}$.

Next, let $r^\star := \lceil k + 2\lambda \rceil$ and $b^\star := q(r^\star)$. If the set $\{b : N_b \geq r^\star\}$ is empty, then $b^\star = B$ by convention and $\hat{b} \leq b^\star$. Otherwise, $N_{b^\star} \geq r^\star \geq k + 2\lambda$. Again using $|Z_{b^\star}| \leq \lambda$ on $\mathcal{E}$, we obtain $\tilde{N}_{b^\star} = N_{b^\star} + Z_{b^\star} \geq N_{b^\star} - |Z_{b^\star}| \geq k + \lambda$. Hence $b^\star \in \{b : \tilde{N}_b \geq k + \lambda\}$. Since $\hat{b}$ is, by definition, the smallest index in the set, it follows that $\hat{b} \leq b^\star$. In both cases, $\hat{b} \leq b^\star = q(\lceil k + 2\lambda \rceil)$.

Combining the two inequalities yields $q(k) \leq \hat{b} \leq q(\lceil k + 2\lambda \rceil)$. Since the grid is increasing in b, this implies $t_{q(k)} \leq t_{\hat{b}} \leq t_{q(\lceil k+2\lambda \rceil)}$. As $\hat{\tau} = t_{\hat{b}}$, the result follows. □

B. Empirical setup

We evaluate on CIFAR-10. The training split is divided into a model-training set and a calibration set, with calibration sizes in $\{1000, 2000, 4000\}$; the test split is used only for empirical evaluation. Models are trained with DP-SGD (Abadi et al., 2016) using Opacus (Yousefpour et al., 2021) for four epochs, batch size 128, training privacy budget $\epsilon_{\text{train}} = 4.0$, and $\delta = 10^{-5}$. We use temperature 2.0, label smoothing 0.5, and calibration failure $\beta = 10^{-3}$. For the card-field sweeps, we vary $\epsilon_{\text{cal}} \in \{2, 4, 8\}$, nominal coverage $\gamma \in \{0.55, 0.65, 0.75, 0.85\}$, calibration size in $\{1000, 2000, 4000\}$, and average over three random seeds. We report empirical coverage, average prediction-set size, certificate width, observed threshold inflation, and classifier accuracy. Accuracy is used only to describe the underlying private classifier; the main conformal utility metric is average prediction-set size. Across the sweeps, the DP-trained classifier has average test accuracy 47.9%, compared with 72.8% for the non-private baseline, so the experiments should be read as a proof of concept for the contract-card diagnostics rather than a claim of state-of-the-art private training. Experiments were run on a single RTX 3080 GPU and take a few hours in total.