PRIVCERT: CERTIFYING STATEMENT SUPPORT UNDER DIFFERENTIAL PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Differentially private (DP) text generation can protect individual records, but privacy alone does not specify what evidence a released statement carries about the underlying data. We identify this as an *evidence gap*: a private report may contain plausible claims without indicating whether they are strongly supported by the private dataset. We introduce PRIVCERT, a framework for privacy-preserving reporting that makes statement support explicit through privacy-preserving certificates and emit-or-abstain decisions. As a canonical instantiation, PRIVCERT-PF (Proposal-and-Filter) separates data-independent candidate discovery from private support certification, emitting only statements whose support passes a private evidence test. We provide theoretical grounding for this framework by characterizing the limits of implicit evidence under DP, deriving a sharp privacy–honesty frontier for single-statement certification, and establishing a worst-case cost for fine-grained multi-statement certification. Experiments on synthetic tasks and TAB, WildChat, and Yelp show that explicit certification maintains low unsupported emission, while free-text DP baselines frequently produce low-support claims under the same declared support semantics. We further show that the PRIVCERT contract can be realized with histogram, sparse-vector, and Gaussian mechanisms, and use DP synthetic data to illustrate an important boundary: support in a private proxy does not automatically certify support in the original data. Together, these results position privacy-preserving reporting as an evidence-design problem: not only how to generate private text, but what a private report can substantiate about its underlying data.

1 INTRODUCTION

Large language models (LLMs) are increasingly used to produce reports and insights from sensitive data, including retrieval-augmented systems over confidential corpora (Lewis et al., 2020), enterprise assistants grounded in private organizational data (Microsoft, 2026), and analytics pipelines over collections of user records (Tamkin et al., 2024; Liu et al., 2025; Cheu et al., 2025). Some such systems additionally run under confidential computing, where intermediate computations and logs may be inaccessible to deployers or auditors (Apple Security Research, 2024; Confidential Computing Consortium, 2023; Google, 2025). In these settings, downstream readers may observe only the released report, making it important to know not only what it says, but how strongly its claims are supported by the underlying data.

Differential privacy (DP) provides a principled way to limit the influence of individual records on released outputs (Dwork et al., 2006; Dwork & Roth, 2014). Recent work extends DP to natural-language generation through private training, synthetic-data generation, and inference-time private decoding (Abadi et al., 2016; Vinod et al., 2025; Xie et al., 2024; Yue et al., 2023). These methods protect individual records, but privacy alone does not specify what evidence a generated statement carries about the dataset.

A DP mechanism may output a plausible statement without indicating whether it reflects a widespread pattern in the private dataset, a rare observation amplified by randomness, or information inherited from the model. For example, from "Users frequently discuss medication side effects," a reader cannot tell how strongly the private corpus supports the claim. We call this mismatch between private text and explicit support evidence the *evidence gap*.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

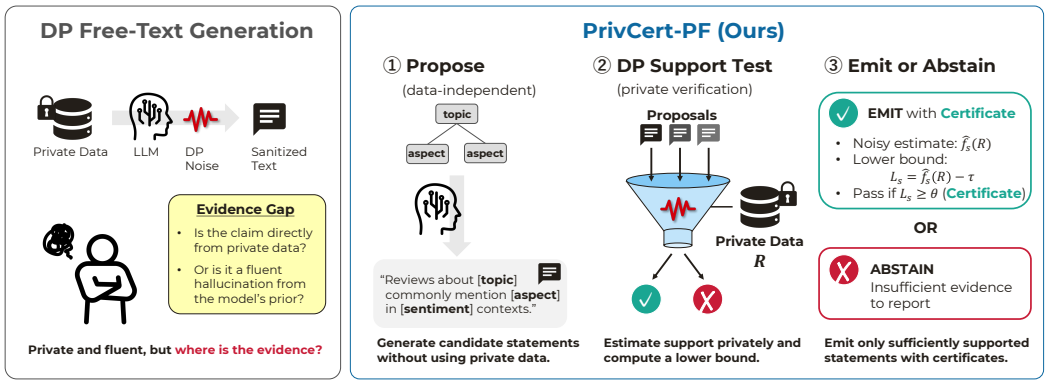

Figure 1: **From private text to private reporting with explicit support evidence.** DP free-text generation can produce fluent private text without exposing the evidential basis of its claims. PrivCert-PF (Proposal-and-Filter), one instantiation of PrivCert, separates candidate discovery from DP support verification and emits only statements that pass a privacy-preserving evidence test.

We formulate privacy-preserving reporting with explicit evidence and introduce PrivCert, a framework in which released statements carry privacy-preserving support evidence and are subject to emit-or-abstain decisions. For a statement s, let $f_s(R) \in [0, 1]$ denote its support in the private dataset R. PrivCert separates candidate discovery, private evidence construction, and release, without prescribing a particular candidate generator or DP primitive.

To ground this framework, we characterize limits on inferring support from ordinary DP outputs and derive a sharp privacy–honesty frontier for single-statement certification. We instantiate the framework with PrivCert-PF (Proposal-and-Filter), which combines data-independent proposals with private support certification, and show that its sharply calibrated half-emission threshold is near-optimal. We further characterize the worst-case cost of fine-grained multi-statement certification and evaluate alternative realizations of the same reporting contract.

Contributions. We make four main contributions:

1. We formulate the *evidence gap* in privacy-preserving text generation and introduce PrivCert, a reporting framework that makes statement support explicit through privacy-preserving certificates and emit-or-abstain decisions (Section 3).
2. We characterize fundamental limits on statement certification under DP: ordinary DP outputs provide only locally limited implicit evidence about support, and explicit certification is constrained by a sharp privacy–honesty frontier (Section 4).
3. We introduce PrivCert-PF, a canonical proposal-and-filter instantiation that separates data-independent candidate discovery from private support certification, and characterize its honesty, coverage, and multi-statement scaling. Under sharp calibration, its single-statement half-emission threshold is within one supporting record of optimal (Section 5).
4. Experiments on synthetic tasks and TAB, WildChat, and Yelp empirically characterize the evidence gap and explicit certification behavior, compare alternative PrivCert realizations, and identify a boundary between direct and proxy-data certification (Section 6).

Together, these results shift privacy-preserving reporting from merely producing private text toward making its evidential basis explicit. They reframe the central question from *how to generate private text* to *what a private report can certify about the data, and at what privacy cost*.

2 Related Work

Differentially private text generation. DP text generation has been studied through private training (Abadi et al., 2016), inference-time decoding (Vinod et al., 2025; Thareja et al., 2026), synthetic text generation (Xie et al., 2024; Yue et al., 2023), and structured-text evaluation (Wang et al., 2025). These works focus on generating useful private text; we instead ask what evidence a released statement provides that the private data support it.

Private aggregate reporting. CLIO summarizes sufficiently large conversation clusters (Tamkin et al., 2024), while Urania provides formal DP guarantees through private clustering, partition selection, and histogram release (Liu et al., 2025). In contrast, we treat candidate statements themselves as objects whose support is privately verified against the dataset.

Faithfulness and abstention. Faithfulness work studies whether generated statements are supported by source material (Ji et al., 2023; Maynez et al., 2020), while selective prediction and conformal methods use abstention under uncertainty (Geifman & El-Yaniv, 2017; Angelopoulos & Bates, 2021). Here, the source records are private, so abstention is driven by a privacy-preserving support test rather than direct verification.

Private query answering, thresholding, and optimal mechanisms. Statement support $f_s(R)$ is a statistical query, connecting our verifier to classical DP query answering (Blum et al., 2008; Hardt & Talwar, 2010; Smith, 2011; Hay et al., 2010), sparse vector techniques (Dwork & Roth, 2014; Lyu et al., 2017), optimal private hypothesis testing (Awan & Slavković, 2018), and staircase mechanisms (Geng & Viswanath, 2016). Our focus is the reporting-specific interface between statements, support certificates, and abstention; our sharp frontier is induced by the dataset-wise one-sided honesty contract as empirical support crosses the reporting threshold.

3 PrivCert: Private Reporting with Explicit Support Evidence

We study privacy-preserving reporting about a private dataset $R = \{r_1, \ldots, r_n\}$ under a reporting task q. Our goal is to formalize not only whether a released report is differentially private, but also what evidence its statements carry about the underlying data.

A mechanism $\mathcal{A}$ is ε-DP if, for each fixed q,

$$\Pr[\mathcal{A}(q, R) \in E] \leq e^{\varepsilon} \Pr[\mathcal{A}(q, R') \in E]$$

for all adjacent datasets $R \sim R'$. By group privacy, datasets at replacement distance at most k satisfy the same inequality with $e^{k\varepsilon}$. This property underlies the lower bounds in Section 4.

Statements and evidence semantics. A statement $s \in \mathcal{S}$ is a semantic claim about the dataset. Its support is represented by $f_s(R) \in [0, 1]$. For a fixed support predicate $\sigma_s(r) \in \{0, 1\}$, we use

$$f_s(R) = \frac{1}{|R|} \sum_{r \in R} \sigma_s(r). \tag{1}$$

This empirical-frequency support has replace-one sensitivity at most $1/|R|$ when σ_s is fixed independently of R. More general semantic matchers may be used, provided their sensitivity and any private adaptation are included in the privacy accounting.

A reporting threshold $\theta \in (0, 1)$ defines insufficient support:

$$f_s(R) < \theta \quad \Longrightarrow \quad s \text{ is unsupported.}$$

Thus f_s specifies the operational notion of evidence, while θ specifies how much evidence is sufficient for reporting. The guarantees below are relative to this specified support semantics; DP does not itself establish matcher-independent semantic truth.

PrivCert interface. A conventional free-text mechanism releases only $Y \sim \mathcal{A}(q, R)$, without explicitly stating how strongly the private dataset supports the claims appearing in Y. A PrivCert mechanism associates each evaluated candidate with (s_j, c_j, a_j), where s_j is a candidate statement, c_j is a privacy-preserving support certificate, and $a_j \in \{\mathrm{emit}, \mathrm{abstain}\}$ is the reporting decision. The released report may expose emitted statements and their certificates, while abstention records that the available evidence was insufficient for release. The interface separates candidate discovery, private evidence construction, and the final release decision; it does not prescribe a particular mechanism for any of these roles.

Framework scope. PrivCert is a reporting contract rather than a new DP primitive. Candidate discovery and evidence construction may use different mechanisms; any private access to R must be included in the privacy accounting, and the resulting certificate must satisfy the stated validity condition. We state the basic results for pure DP for clarity, while the framework also allows

approximate-DP mechanisms. PRIVCERT-PF in Section 5 is one direct-certification instantiation of this contract.

For the one-sided certificates studied here, c_s is valid at failure probability β if

$$\Pr[c_s > f_s(R)] \le \beta.$$

Equivalently, except with probability β, the true dataset support is at least as large as the reported certificate c_s. Under the threshold reporting rule used by the instantiations studied here, $a_s = \text{emit} \implies c_s \ge \theta$, validity implies

$$f_s(R) < \theta \quad \implies \quad \Pr[a_s = \text{emit}] \le \beta. \tag{2}$$

This is the basic evidential contract (PRIVCERT contract): an unsupported statement is emitted only with controlled probability. For multi-statement reports, the evidential error scope must also be specified: per-candidate validity does not by itself control the probability that any unsupported statement appears in the complete report.

4 PRIVACY LIMITS AND OPTIMAL STATEMENT CERTIFICATION

We study two limits on statement evidence under differential privacy. First, we ask how reliably support can be inferred from a single DP output. Second, under the explicit one-sided honesty contract of Section 3, we characterize the largest possible emission probability of any private certifier. The latter yields a sharp frontier used to establish the near-optimality of PRIVCERT-PF in Section 5. Proofs are given in Appendix A.

4.1 LIMITS OF IMPLICIT EVIDENCE

Fix a statement s and datasets R_0, R_1 with $d(R_0, R_1) \le k$. Let $P_i := \mathcal{L}(\mathcal{A}(q, R_i))$, $i \in \{0, 1\}$.

Theorem 1 (Local indistinguishability under DP)**.** *For any ε-DP mechanism $\mathcal{A}$,*

$$d_{\mathrm{TV}}(P_0, P_1) \le \frac{e^{k\varepsilon} - 1}{e^{k\varepsilon} + 1}.$$

Consequently, every possibly randomized binary test φ satisfies

$$\mathbb{E}_{P_0}[\varphi(Y)] + \mathbb{E}_{P_1}[1 - \varphi(Y)] \ge \frac{2}{e^{k\varepsilon} + 1}.$$

This is the standard binary-testing consequence of group privacy (Kairouz et al., 2015), applied here to statement support. In particular, the occurrence of a statement in a DP output cannot by itself serve as a calibrated support certificate.

The same argument limits support estimation. Writing $a_i := f_s(R_i)$ and $\Delta := |a_1 - a_0|$, every estimator $\hat{f}_s : \mathcal{Y} \to [0, 1]$ satisfies

$$\max_{i \in \{0,1\}} \mathbb{E}_{Y \sim P_i}\big[|\hat{f}_s(Y) - a_i|\big] \ge \frac{\Delta}{2(e^{k\varepsilon} + 1)}.$$

For the empirical-frequency support in Equation 1, changing support by Δ requires at least $\lceil n\Delta \rceil$ record replacements, so these bounds are local and become weak when $k\varepsilon$ is large. They do not imply that DP hides arbitrarily large support differences; rather, a single DP output cannot provide arbitrarily reliable implicit evidence on nearby datasets.

4.2 SHARP SINGLE-STATEMENT CERTIFICATION FRONTIER

We now ask how quickly any private certifier can increase its probability of releasing a statement as its support rises above the reporting threshold.

For a fixed statement s, let $N_s(R) := \sum_{r \in R} \sigma_s(r)$, $t_\theta := \lceil n\theta \rceil$, so that $t_\theta - 1$ is the largest unsupported support count. Define

$$d_s(R) := \max\{N_s(R) - t_\theta + 1, 0\},$$

the number of supporting-record steps from this boundary.

Theorem 2 (Sharp single-statement emission frontier)**.** *Assume $0 < \beta < 1/2$ and that both supporting and non-supporting records exist. Let $\pi_0 := \beta$, $\pi_{k+1} := \min\{e^{\varepsilon}\pi_k,\, 1 - e^{-\varepsilon}(1-\pi_k)\}$. For any ε-DP mechanism satisfying $f_s(R) < \theta \Longrightarrow \Pr[a_s = \text{emit} \mid R] \le \beta$, the emission probability satisfies*

$$\Pr[a_s = \text{emit} \mid R] \le \pi_{d_s(R)}$$

for every dataset R. Moreover, the bound is pointwise tight: an ε-DP, β-honest count-based mechanism attains $\pi_{d_s(R)}$ at every support count.

The theorem gives an exact upper envelope on how quickly release can become more likely as support increases. Since $\pi_k \le \beta e^{k\varepsilon}$, any certifier with emission probability at least $1/2$ must satisfy

$$d_s(R) \ge \left\lceil \frac{1}{\varepsilon} \log \frac{1}{2\beta} \right\rceil.$$

The optimal count-based mechanism attains the binary release frontier but provides only a threshold-level certificate. Section 5 shows that PRIVCERT-PF reaches this fundamental margin essentially optimally while additionally providing a quantitative lower certificate on statement support.

5 A CANONICAL PRIVCERT INSTANTIATION: PROPOSAL-AND-FILTER

We instantiate PRIVCERT with PRIVCERT-*PF (Proposal-and-Filter)*, which separates data-independent candidate discovery from private support certification and emits only statements whose lower support certificates clear the reporting threshold. The construction is deliberately simple, making its privacy, evidential validity, and coverage behavior explicit. Section 6 evaluates alternative PRIVCERT instantiations with different sensitivity and selection structures. Proofs and pseudocode are given in Appendix B.

5.1 PROPOSAL-AND-FILTER

Given a reporting task q, let Π_q be a proposal distribution independent of the private dataset R, and sample $s_1, \ldots, s_m \overset{\text{i.i.d.}}{\sim} \Pi_q$. For each candidate s_j, PRIVCERT-PF computes

$$\widehat{f}_{s_j}(R) = f_{s_j}(R) + Z_j, \qquad L_{s_j} = \widehat{f}_{s_j}(R) - \tau,$$

and emits s_j iff $L_{s_j} \ge \theta$. Our primary instantiation uses independent Laplace noise with equal allocation of the total privacy budget; Section 5.2 analyzes its calibration, and Appendix C.2 gives the experimental accounting.

Candidate generation is therefore a discovery step rather than a source of evidential validity: proposals need not themselves be correct, but every released statement must pass the private support test. Conversely, a supported statement that is never proposed cannot be certified. Any final natural-language rendering is post-processing, but must preserve the semantics of the certified statement; otherwise the certificate applies only to the original candidate.

5.2 PRIVACY, HONESTY, AND NEAR-OPTIMALITY

If support query j is ε_j-DP, then data-independent proposal, sequential composition, and post-processing make PRIVCERT-PF $\sum_j \varepsilon_j$-DP.

Theorem 3 (Certificate validity and per-candidate one-sided honesty)**.** *Fix an evaluated statement s. Suppose $\widehat{f}_s(R) = f_s(R) + Z$, $\Pr[Z \ge \tau] \le \beta$, $L_s = \widehat{f}_s(R) - \tau$. Then*

$$\Pr[L_s > f_s(R)] \le \beta.$$

Consequently, if s is emitted only when $L_s \ge \theta$, then $f_s(R) < \theta \Longrightarrow \Pr[\text{Emit}_s] \le \beta$.

The guarantee remains valid after candidate selection when the certification noise is fresh and independent of the selection event; any private selection must additionally be included in the privacy accounting. For m evaluated candidates, a union bound gives a report-level failure probability at most $m\beta$; Appendix C.7 evaluates calibration to a fixed whole-report target.

Near-optimality. The sharp frontier in Theorem 2 also characterizes how close PRIVCERT-PF is to the best possible single-statement certifier. For a statement with privacy budget ε_s, consider

$$Z \sim \mathrm{Lap}\left(\frac{1}{n\varepsilon_s}\right), \qquad \tau = \frac{1}{n\varepsilon_s}\log\frac{1}{2\beta},$$

the tight symmetric Laplace calibration for the two one-sided tail conditions.

Corollary 1 (Near-optimal sharply calibrated half-emission threshold)**.** *Let $A := \frac{1}{\varepsilon_s}\log\frac{1}{2\beta}$, $k^\star := \lceil A \rceil$, and assume $\lceil n\theta + A \rceil \le n$, equivalently, $\theta + \tau \le 1$ for the sharp Laplace margin $\tau = A/n$.*

For any ε_s-DP, β-honest single-statement certifier that reaches emission probability at least $1/2$, the first such distance from the largest unsupported dataset is at least $k^\star$. Sharply calibrated Laplace PRIVCERT*-PF first reaches emission probability at least $1/2$ at distance either $k^\star$ or $k^\star + 1$. Hence its half-emission threshold is within one supporting record of optimal.*

Thus, under sharp calibration, little room remains to improve the half-emission threshold once the per-statement privacy budget is fixed. The optimal staircase mechanism of Theorem 2 achieves the exact binary release frontier but exposes only a threshold-level certificate; PRIVCERT-PF additionally provides a quantitative lower certificate on statement support.

5.3 DISCOVERY AND COVERAGE

Certification alone does not determine whether useful supported statements are discovered. For $a \in \mathbb{R}$, define

$$p_a(R) := \Pr_{s\sim\Pi_q}[f_s(R) \ge a],$$

which measures how much proposal mass lies above a given support level.

Theorem 4 (EmitRate trade-off)**.** *Let $s_1, \ldots, s_m \overset{\text{i.i.d.}}{\sim} \Pi_q$, and suppose $\hat{f}_{s_j}(R) = f_{s_j}(R) + Z_j$, where Z_j is independent of s_j and satisfies $\Pr[Z_j \ge \tau] \le \beta$ and $\Pr[Z_j \le -\tau] \le \beta$. Let*

$$O := \{j : \hat{f}_{s_j}(R) - \tau \ge \theta\}, \qquad \mathrm{EmitRate} := \mathbb{E}\left[\frac{|O|}{m}\right].$$

Then

$$(1-\beta)p_{\theta+2\tau}(R) \le \mathrm{EmitRate} \le p_\theta(R) + \beta\big(1 - p_\theta(R)\big) \le p_\theta(R) + \beta.$$

The theorem separates discovery from verification. The verifier controls unsupported emission, while coverage depends on proposal mass in regions that can be certified under the available privacy budget. In particular, candidates with $f_s(R) \ge \theta + 2\tau$ are emitted with probability at least $1 - \beta$.

Tag-template proposer. Our primary proposer is designed around three principles: *data independence*, so discovery incurs no privacy cost; *matcher alignment*, so proposed statements can be evaluated reliably under the fixed support semantics; and *threshold coverage*, so proposal mass is placed on statements likely to have sufficient support. We instantiate these principles using a public schema over domain-relevant dimensions such as topic, aspect, and sentiment. A sampled tuple is rendered through a fixed template, for example,

> "Reviews about `{topic}` commonly mention `{aspect}` in `{sentiment}` contexts."

The schema, values, and templates are fixed before observing R. Controlled surface forms make matcher behavior easier to validate on public or otherwise non-sensitive examples, while schema granularity trades specificity against the probability of clearing the reporting threshold. A public-LLM proposer provides another data-independent discovery mechanism; we compare the two in Appendix C.4.

5.4 THE COST OF FINE-GRAINED CERTIFICATION

The single-statement analysis assumes that a privacy budget has already been assigned to one statement. With m statements sharing a total budget $\varepsilon_{\mathrm{tot}}$, one-sided honesty alone is insufficient to express useful coverage, since always abstaining is trivially honest. We therefore call a mechanism (w, β)-fine-grained if, for every statement s_j,

$$f_{s_j}(R) < \theta \Longrightarrow \Pr[j \in O] \le \beta, \qquad f_{s_j}(R) \ge \theta + w \Longrightarrow \Pr[j \in O] \ge 1 - \beta.$$

Theorem 5 (Cost of fine-grained certification)**.** *Consider a worst-case regime in which the record universe can realize every binary support pattern over m statements, and assume $\lceil n\theta \rceil + \lceil nw \rceil \leq n$. For fixed $0 < \beta < 1/8$ and $\varepsilon_{\text{tot}} = o(m)$, any pure ε_{tot}-DP mechanism satisfying (w, β)-fine-grained certification must have*

$$w = \Omega\left(\frac{m}{n\varepsilon_{\text{tot}}}\right).$$

The finite-sample bound and packing proof are given in Appendix B. The richness assumption is a worst-case condition rather than a claim about every semantic support model.

Under equal allocation, sharply calibrated Laplace PRIVCERT-PF satisfies fine-grained certification with $w_{\text{PF}} = \frac{2m}{n\varepsilon_{\text{tot}}} \log \frac{1}{2\beta}$, matching the lower bound's dependence on $m/(n\varepsilon_{\text{tot}})$ in this regime for fixed β.

Thus, in this regime, the linear scaling is not merely an artifact of Laplace noise or equal budget allocation: in the worst-case overlapping-support regime, it is the cost of resolving many statement supports individually. Other PRIVCERT instantiations can improve this trade-off when additional structure, joint sensitivity, or selective release can be exploited.

6 EMPIRICAL EVALUATION

We evaluate whether the PRIVCERT framework developed above is reflected in practice. Our experiments ask whether (i) free-text DP baselines emit claims with low support, (ii) PRIVCERT-PF controls unsupported emission, (iii) its EmitRate follows Theorem 4, and (iv) alternative PRIVCERT instantiations realize the same reporting contract under different trade-offs.

Metrics. For PRIVCERT-PF, we report UnsupportedEmit, the empirical per-candidate failure rate controlled by Theorem 3, and EmitRate, the coverage quantity characterized by Theorem 4. We additionally report FalseEmission and MeanSupport as descriptive diagnostics; full definitions are given in Appendix C.

Additional analyses of proposer design, semantic-matcher robustness, and family-wise guarantees are reported in Appendices C.4, C.6, and C.7, respectively.

6.1 SYNTHETIC VALIDATION

We first isolate the PRIVCERT-PF evidence mechanism from proposer and classifier effects using a finite statement universe with oracle support values. Candidates are sampled uniformly, and we sweep $\varepsilon_{\text{tot}} \in \{1, 4, 10, 20\}$, $\theta \in \{0.05, 0.1, 0.2\}$, $m \in \{50, 200, 1000\}$, and $\beta \in \{0.01, 0.05, 0.1\}$ over five seeds. The results are consistent with the one-sided calibration and Theorem 4 across the full grid. UnsupportedEmit remains low under the prescribed one-sided calibration (maximum 0.038, mean 0.002), and empirical EmitRate lies within the interval predicted by Theorem 4 in all tested cells up to sampling slack. The sweep also illustrates the effect of m under a fixed total privacy budget: increasing m evaluates more candidates in absolute number but reduces the privacy budget available to each support query, increasing the certification margin. Because EmitRate is a fraction, increasing m does not itself increase the proposal-induced support mass $p_a(R)$. Full synthetic settings and diagnostics are reported in Appendix C.1.

Takeaway. Across the synthetic sweep, explicit certification reliably suppresses unsupported statements, while its coverage changes as predicted with the available privacy budget.

6.2 REALISTIC CORPUS EXPERIMENTS

We next evaluate whether PRIVCERT-PF controls unsupported emission in realistic text-reporting settings, and whether the evidence gap also appears in free-text DP outputs. We use TAB (legal text) (Pilán et al., 2022), WildChat (conversations) (Zhao et al., 2024), and Yelp (reviews) (Zhang et al., 2015).

Support and proposer. We use a frozen public NLI cross-encoder to define the record-level support predicate and the tag-template proposer from Section 5.3. The matcher is fixed independently of R, giving sensitivity at most $1/|R|$. Full support-model and calibration details are given in Appendix C.2.

Table 1: **FalseEmission on realistic datasets under the declared support semantics.** For free-text methods this is a descriptive diagnostic, not a theorem-backed error rate. Lower is better.

Dataset	Unfiltered	Doc-DP	INVISIBLEINK (TinyLlama)	INVISIBLEINK (Qwen)	PRIVCERT-PF
TAB	1.000	0.894	0.958	0.994	**0.010**
WildChat	0.947	0.729	0.647	0.619	**0.014**
Yelp	0.943	0.803	0.877	0.856	**0.000**

Table 2: **Privacy-budget sweep on WildChat.** Larger budgets improve EmitRate while preserving one-sided honesty. Intervals use the tightened upper bound $p_\theta + \beta(1 - p_\theta)$.

ε_{tot}	EmitRate ↑	UnsupEmit ↓	FalseEmit ↓	MeanSupport ↑	Interval	In?
1	0.051	0.0217	0.295	0.275	[0.000, 0.335]	✓
4	0.180	0.0085	0.033	0.327	[0.135, 0.335]	✓
10	0.220	0.0046	0.014	0.288	[0.162, 0.335]	✓
20	0.274	0.0014	0.004	0.250	[0.189, 0.335]	✓

Baselines. We compare against no-data unfiltered generation, an aggregate-then-render document-level DP baseline, and token-level DP decoding with INVISIBLEINK (Vinod et al., 2025), all at the common target $(\varepsilon_{\text{tot}}, \delta) = (10, 10^{-5})$. Details are given in Appendix C.2.

Table 1 summarizes the main results.

Free-text DP outputs often contain low-support claims under the declared support semantics. FalseEmission remains high for both document-level and token-level DP baselines. For INVISIBLEINK, it ranges from 0.619 to 0.994 across datasets and model backbones. Thus, privacy-preserving generation alone does not make statement occurrence an explicit support certificate. Because the methods differ in output structure and candidate-generation procedure, Table 1 should be read as a diagnostic of the evidence gap under a common support measure, rather than as a causal estimate of the effect of certification or a general ranking of semantic quality.

PRIVCERT-PF realizes the PRIVCERT contract. Across the three datasets, UnsupportedEmit lies in $[0.000, 0.008]$, below the target $\beta = 0.05$. This is the per-candidate quantity controlled by Theorem 3. The same behavior holds across proposer ablations (Appendix C.4). We further evaluate PRIVCERT-PF with a frozen Qwen3-8B judge as an alternative support model in Appendix C.6.

EmitRate is consistent with the predicted trade-off. The observed aggregate EmitRates and all 15 per-seed realizations fall within the interval predicted for the expectation by Theorem 4. For the aggregate results, TAB: $0.404 \in [0.356, 0.532]$, WildChat: $0.220 \in [0.162, 0.335]$, and Yelp: $0.749 \in [0.620, 0.809]$. These finite-run checks are empirical consistency tests rather than additional per-run guarantees.

Privacy-budget sweep. We evaluate the coverage–noise trade-off on WildChat while keeping the tag-template proposer, support model, threshold, and per-candidate honesty target fixed. We vary $\varepsilon_{\text{tot}} \in \{1, 4, 10, 20\}$ with $\theta = 0.05$ and $\beta = 0.05$. As ε_{tot} increases, support estimates become less noisy and EmitRate rises from 0.051 to 0.274. UnsupportedEmit remains below $\beta = 0.05$ throughout. The results are consistent with the inverse-budget dependence of the certification margin characterized in Section 5.4.

Whole-report control. The guarantee above applies to each candidate separately: an unsupported candidate is emitted with probability at most β. When many candidates are evaluated, these risks can accumulate, so this does not bound the probability that the final report contains any unsupported statement. To obtain such a report-level guarantee, we allocate a total failure budget β_{tot} across candidates so that $\Pr[\exists s \in O : \ f_s(R) < \theta] \leq \beta_{\text{tot}}$. With $\beta_{\text{tot}} = 0.05$, EmitRate decreases from 0.404 to 0.357 on TAB, 0.220 to 0.169 on WildChat, and 0.749 to 0.565 on Yelp (Appendix C.7).

Takeaway. On realistic corpora, PRIVCERT-PF maintains low unsupported emission and the predicted coverage behavior, while free-text DP outputs frequently contain low-support claims under the same declared evidence semantics.

Table 3: **Matched comparisons of PRIVCERT instantiations.** Entries show EmitRate. All UnsupportedEmit values remain below $\beta = 0.05$. Comparisons are matched within each candidate space.

	Sampled $m = 200$		Exact-unique		
Dataset	PRIVCERT-Hist	PRIVCERT-PF	PRIVCERT-SVT	PRIVCERT-PF	PRIVCERT-PF (Gaus)
TAB	0.199	0.404	0.400	0.422	0.444
WildChat	0.183	0.220	0.151	0.151	0.162
Yelp	0.390	0.749	0.796	0.816	0.833

6.3 ALTERNATIVE PRIVCERT INSTANTIATIONS

We next evaluate alternative instantiations of the PRIVCERT framework. PRIVCERT-Hist uses a contribution-bounded joint support release, PRIVCERT-SVT combines private sparse selection with subsequent certification, and a Gaussian variant of PRIVCERT-PF replaces the pure-DP Laplace mechanism with an analytic Gaussian mechanism under (ε, δ)-DP. Table 3 reports EmitRate; all corresponding UnsupportedEmit values remain below the target $\beta = 0.05$. Full mechanism definitions and privacy accounting are given in Appendix C.3.

PRIVCERT-Hist changes sensitivity through contribution bounding, while PRIVCERT-SVT privately selects a bounded candidate set before certification, so Theorem 4 does not apply unchanged. The Gaussian variant changes only the DP backend. Together, these matched comparisons show that the PRIVCERT contract is compatible with different evidence constructions and privacy mechanisms.

A boundary case: private proxies. An Aug-PE (Xie et al., 2024)-style DP synthetic-data backend illustrates a boundary of the PRIVCERT contract: $f_s(\widetilde{R}) \geq \theta$ does not imply $f_s(R) \geq \theta$. At $\theta = 0.10$, the synthetic backend attains similar EmitRate (0.676 versus 0.695 for direct certification) but an original-data UnsupportedEmit of 0.092, compared with 0.001 for PRIVCERT-PF. Original-data certification therefore requires an additional fidelity guarantee (Appendix C.5).

Takeaway. The PRIVCERT contract is not tied to a particular DP mechanism, but what is certified still depends on the reference data: support in a private proxy does not by itself certify support in the original dataset.

7 DISCUSSION AND LIMITATIONS

What PRIVCERT certifies. PRIVCERT guarantees are relative to the declared reference data and support function f_s, not matcher-independent semantic truth. Different support models can preserve mechanism-level honesty while changing coverage and certified content (Appendix C.6). Support measured on DP synthetic data certifies only that proxy absent an additional fidelity guarantee, and statements never proposed cannot be certified.

Honesty does not imply completeness. Privacy noise, conservative certification, and limited discovery may cause abstention, and statement-level certificates do not guarantee logical consistency of the complete report.

8 CONCLUSION

We formulated the *evidence gap* in privacy-preserving text generation: differential privacy protects individual records but does not specify what evidence a released statement carries about the underlying data. We introduced PRIVCERT, a reporting framework that makes statement support explicit through privacy-preserving certificates and emit-or-abstain decisions. We provided theoretical grounding for this framework by characterizing fundamental privacy–honesty limits, showing that sharply calibrated PRIVCERT-PF operates near the single-statement optimum, and establishing a worst-case cost for fine-grained multi-statement certification. Experiments across multiple datasets and DP mechanisms empirically characterize the evidence gap and demonstrate alternative realizations of the reporting contract while clarifying its scope and boundaries. Together, these results recast privacy-preserving reporting as an *evidence-design problem*: not only *how to generate private text*, but *what a private report can certify about the data, and at what privacy cost*.

AI USE STATEMENT

Generative AI tools were used for manuscript drafting, language editing, organization, and proof-reading, and for feedback on experimental methodology, theoretical presentation, comparisons, and interpretation of results. All reported experimental measurements were obtained from executed experiments or deterministic analyses of their outputs. The authors reviewed and verified all AI-assisted text, technical claims, proofs, experimental results, analyses, and citations, and take responsibility for the final content.

ETHICS STATEMENT

This work studies privacy-preserving reporting from sensitive text collections. The experiments use public datasets and public language models, and do not involve new human-subject recruitment or interaction. Some evaluated datasets contain user-generated text; our experiments use them only to study aggregate reporting behavior and do not attempt to identify individuals, infer user-level attributes, or release record-level information.

The proposed framework is designed to make explicit what support a released statement has in a reference dataset while limiting the influence of individual records through differential privacy. Its guarantees are relative to a declared support function and should not be interpreted as certifying semantic truth independently of that definition. Likewise, privacy and support certification do not eliminate all risks associated with deploying language models on sensitive data, and application specific review may still be required.

REPRODUCIBILITY STATEMENT

We report the privacy parameters, support semantics, proposal mechanisms, thresholds, random seeds, and evaluation metrics used in the experiments. Synthetic experiments use repeated seeded trials over the stated parameter grid, and realistic PRIVCERT-PF experiments use repeated candidate pools and recorded candidate-level support values, noise draws, certificates, and release decisions. Free-text baselines are evaluated using a fixed post-processing procedure for claim extraction and the same declared support function used for the main diagnostics.

The appendices document the support model, privacy accounting, proposal and calibration procedures, baseline implementations, experimental repetition protocols, theorem checks, alternative PRIVCERT instantiations, and robustness analyses. Where experiments use different repetition units or candidate spaces, we report those differences explicitly and avoid treating unmatched settings as direct statistical comparisons.

REFERENCES

Martín 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 (CCS)*, pp. 308–318. ACM, 2016. doi: 10.1145/2976749.2978318.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. https://arxiv.org/abs/2107.07511.

Apple Security Research. Private cloud compute security guide. https://security.apple.com/documentation/private-cloud-compute/, 2024. Accessed 2026.

Jordan Awan and Aleksandra Slavković. Differentially private uniformly most powerful tests for binomial data. In *Advances in Neural Information Processing Systems*, volume 31, pp. 4212–4222, 2018.

Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 609–618. ACM, 2008. doi: 10.1145/1374376.1374464.

Albert Cheu, Artem Lagzdin, Brett McLarnon, Daniel Ramage, Katharine Daly, Marco Gruteser, Peter Kairouz, Rakshita Tandon, Stanislav Chiknavaryan, Timon Van Overveldt, et al. Toward provably private analytics and insights into genai use. *arXiv preprint arXiv:2510.21684*, 2025.

Confidential Computing Consortium. A technical analysis of confidential computing. Technical report, Confidential Computing Consortium, 2023. https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/CCC-A-Technical-Analysis-of-Confidential-Computing-v1.3_unlocked.pdf.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9 of *Foundations and Trends in Theoretical Computer Science*. Now Publishers, 2014. doi: 10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference (TCC)*, volume 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer, 2006. doi: 10.1007/11681878_14.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Quan Geng and Pramod Viswanath. The optimal noise-adding mechanism in differential privacy. *IEEE Transactions on Information Theory*, 62(2):925–951, 2016. doi: 10.1109/TIT.2015.2504967.

Google. Private ai compute in the cloud. https://services.google.com/fh/files/misc/private_ai_compute_technical_brief.pdf, 2025. Accessed 2026.

Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)*, pp. 705–714. ACM, 2010. doi: 10.1145/1806689.1806786.

Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*, 3(1–2):1021–1032, 2010. doi: 10.14778/1920841.1920970.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1376–1385. PMLR, 2015.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. https://arxiv.org/abs/2005.11401.

Daogao Liu, Edith Cohen, Badih Ghazi, Peter Kairouz, Pritish Kamath, Alexander Knop, Ravi Kumar, Pasin Manurangsi, Adam Sealfon, Da Yu, and Chiyuan Zhang. Urania: Differentially private insights into AI use. In *Conference on Language Modeling (COLM)*, 2025. https://arxiv.org/abs/2506.04681.

Min Lyu, Dong Su, and Ninghui Li. Understanding the sparse vector technique for differential privacy. *Proceedings of the VLDB Endowment*, 10(6):637–648, 2017.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1906–1919. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.173.

Microsoft. Data, privacy, and security for microsoft 365 copilot. https://learn.microsoft.com/en-us/microsoft-365/copilot/microsoft-365-copilot-privacy, 2026. Accessed 2026.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101, 2022.

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing (STOC)*, pp. 813–822. ACM, 2011. doi: 10.1145/1993636.1993743.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, Michael Stern, Brian Clarke, Landon Goldberg, Theodore R. Sumers, Jared Mueller, William McEachen, Wes Mitchell, Shan Carter, Jack Clark, Jared Kaplan, and Deep Ganguli. Clio: Privacy-preserving insights into real-world AI use. Technical report, Anthropic, 2024. https://arxiv.org/abs/2412.13678.

Rushil Thareja, Preslav Nakov, Praneeth Vepakomma, and Nils Lukas. Dp-fusion: Token-level differentially private inference for large language models. In *International Conference on Learning Representations*, 2026. https://openreview.net/forum?id=WLK37mn0El.

Vishnu Vinod, Krishna Pillutla, and Abhradeep Guha Thakurta. Invisibleink: High-utility and low-cost text generation with differential privacy. In *Advances in Neural Information Processing Systems*, 2025.

Shuaiqi Wang, Vikas Raunak, Arturs Backurs, Victor Reis, Pei Zhou, Sihao Chen, Longqi Yang, Zinan Lin, Sergey Yekhanin, and Giulia Fanti. Struct-bench: A benchmark for differentially private structured text generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, et al. Differentially private synthetic data via foundation model apis 2: Text. In *International Conference on Machine Learning*, pp. 54531–54560. PMLR, 2024.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1321–1342, 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. The Yelp Reviews Full dataset, derived from the Yelp Dataset Challenge 2015, was first introduced as a text classification benchmark in this paper. https://arxiv.org/abs/1509.01626.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT interaction logs in the wild. In *International Conference on Learning Representations (ICLR)*, 2024. https://openreview.net/forum?id=Bl8u7ZRlbM.

APPENDIX

A PROOFS FOR SECTION 4

A.1 PROOF OF THEOREM 1 AND SUPPORT-ESTIMATION CONSEQUENCE

Let $P_i := \mathcal{L}(\mathcal{A}(q, R_i))$, $t := e^{k\varepsilon}$. By group privacy, for every measurable event E,

$$P_0(E) \le tP_1(E), \qquad P_1(E) \le tP_0(E).$$

To bound total variation, fix an event B with $P_0(B) \ge P_1(B)$, and write $p := P_0(B)$, $q := P_1(B)$, and $\Delta := p - q$. Applying the two group-privacy inequalities to B and B^c gives

$$\Delta \le (t-1)q, \qquad \Delta \le \frac{t-1}{t}(1-q).$$

The minimum of these two bounds is maximized at $q = 1/(t+1)$, hence

$$d_{\mathrm{TV}}(P_0, P_1) \le \frac{t-1}{t+1} = \frac{e^{k\varepsilon}-1}{e^{k\varepsilon}+1}.$$

The standard binary-testing identity then gives, for every possibly randomized test $\varphi : \mathcal{Y} \to [0,1]$,

$$\mathbb{E}_{P_0}[\varphi(Y)] + \mathbb{E}_{P_1}[1-\varphi(Y)] \ge 1 - d_{\mathrm{TV}}(P_0, P_1) \ge \frac{2}{e^{k\varepsilon}+1},$$

which proves Theorem 1.

For the support-estimation consequence, let $a_i := f_s(R_i)$ and $\Delta_s := |a_1 - a_0|$, and assume without loss of generality that $a_1 > a_0$. Given any estimator $\hat{f}_s : \mathcal{Y} \to [0,1]$, define

$$\varphi(Y) := \mathbf{1}\left\{\hat{f}_s(Y) \ge \frac{a_0 + a_1}{2}\right\}.$$

If $\varphi(Y) = 1$ under P_0, or $\varphi(Y) = 0$ under P_1, the estimation error is at least $\Delta_s/2$. Therefore,

$$\begin{aligned}&\mathbb{E}_{P_0}\left[|\hat{f}_s(Y) - a_0|\right] + \mathbb{E}_{P_1}\left[|\hat{f}_s(Y) - a_1|\right] \\ &\quad\ge \frac{\Delta_s}{2}\left(\mathbb{E}_{P_0}[\varphi(Y)] + \mathbb{E}_{P_1}[1-\varphi(Y)]\right).\end{aligned}$$

Since the maximum of two nonnegative quantities is at least half their sum, Theorem 1 implies

$$\max_{i\in\{0,1\}} \mathbb{E}_{Y\sim P_i}\left[|\hat{f}_s(Y) - a_i|\right] \ge \frac{\Delta_s}{2(e^{k\varepsilon}+1)}.$$

A.2 PROOF OF THEOREM 2

Fix a statement s, and write

$$p_{\mathcal{A}}(R) := \Pr[a_s = \mathrm{emit} \mid R].$$

For adjacent datasets $R \sim R'$, differential privacy applied to the emission event and to the complementary abstention event gives

$$p_{\mathcal{A}}(R') \le e^{\varepsilon} p_{\mathcal{A}}(R),$$

and

$$p_{\mathcal{A}}(R') \le 1 - e^{-\varepsilon}\big(1 - p_{\mathcal{A}}(R)\big).$$

Thus, if one additional supporting record is introduced,

$$p_{\mathcal{A}}(R') \le g_{\varepsilon}(p_{\mathcal{A}}(R)),$$

where

$$g_{\varepsilon}(p) := \min\left\{e^{\varepsilon}p,\ 1 - e^{-\varepsilon}(1-p)\right\}.$$

The map g_{ε} is nondecreasing.

Upper bound. Let $\pi_0 := \beta$, $\pi_{k+1} := g_\varepsilon(\pi_k)$. We prove by induction on $d_s(R)$ that

$$p_{\mathcal{A}}(R) \le \pi_{d_s(R)}.$$

For $d_s(R) = 0$, the statement is unsupported, so the honesty condition gives

$$p_{\mathcal{A}}(R) \le \beta = \pi_0.$$

For $d_s(R) = k + 1$, replace one supporting record by a fixed non-supporting record to obtain an adjacent dataset R^- with $d_s(R^-) = k$. Then

$$p_{\mathcal{A}}(R) \le g_\varepsilon\big(p_{\mathcal{A}}(R^-)\big) \le g_\varepsilon(\pi_k) = \pi_{k+1},$$

where the second inequality uses the induction hypothesis and monotonicity of g_ε.

Tightness. Consider the count-based mechanism $\mathcal{A}^\star$ that emits s with probability

$$p^\star(R) := \pi_{d_s(R)}$$

and otherwise abstains. Adjacent datasets have d_s-values differing by at most one. For consecutive k and $k + 1$, the recurrence gives

$$\pi_{k+1} \le e^\varepsilon \pi_k, \qquad 1 - \pi_k \le e^\varepsilon (1 - \pi_{k+1}).$$

Because $\pi_{k+1} \ge \pi_k$, the reverse-direction DP inequalities hold as well. Hence the binary mechanism is ε-DP.

For unsupported datasets $d_s(R) = 0$, so $p^\star(R) = \beta$, establishing β-honesty. Thus the frontier is attained simultaneously at every support count. Attaching $c_s = \theta$ upon emission and $c_s = 0$ otherwise also satisfies the PRIVCERT certificate-validity condition.

Half-emission consequence. Since $g_\varepsilon(p) \le e^\varepsilon p$, we have $\pi_k \le \beta e^{k\varepsilon}$. Therefore, emission probability at least $1/2$ requires

$$d_s(R) \ge \left\lceil \frac{1}{\varepsilon} \log \frac{1}{2\beta} \right\rceil,$$

which is the margin lower bound used in Corollary 1.

A.3 Extension to Approximate Differential Privacy

The local indistinguishability results in Section 4.1 extend to (ε, δ)-DP. For datasets R_0, R_1 with $d(R_0, R_1) \le k$, define

$$\delta_k := \delta \sum_{\ell=0}^{k-1} e^{\ell\varepsilon}, \qquad t := e^{k\varepsilon}.$$

Approximate group privacy gives

$$P_0(E) \le tP_1(E) + \delta_k, \qquad P_1(E) \le tP_0(E) + \delta_k$$

for every measurable event E.

For an event B with $P_0(B) \ge P_1(B)$, writing $q := P_1(B)$ and $\Delta := P_0(B) - P_1(B)$, these inequalities imply

$$\Delta \le (t-1)q + \delta_k, \qquad \Delta \le \frac{t-1}{t}(1-q) + \frac{\delta_k}{t}.$$

Maximizing the smaller of the two bounds yields

$$d_{\mathrm{TV}}(P_0, P_1) \le \min\left\{1,\ \frac{t - 1 + 2\delta_k}{t+1}\right\}.$$

Consequently, every possibly randomized binary test satisfies

$$\mathbb{E}_{P_0}[\varphi(Y)] + \mathbb{E}_{P_1}[1 - \varphi(Y)] \ge \left[\frac{2(1-\delta_k)}{e^{k\varepsilon} + 1}\right]_+,$$

where $[x]_+ := \max\{x, 0\}$.

Applying the same two-point reduction as above, with $a_i := f_s(R_i)$ and $\Delta_s := |a_1 - a_0|$, gives

$$\max_{i \in \{0,1\}} \mathbb{E}_{Y \sim P_i}\left[|\hat{f}_s(Y) - a_i|\right] \ge \left[\frac{\Delta_s(1-\delta_k)}{2(e^{k\varepsilon}+1)}\right]_+.$$

Setting $\delta = 0$ recovers the pure-DP bounds. As in the pure-DP case, these local bounds may become weak when $k\varepsilon$ or the accumulated slack δ_k is large.

B Additional Details and Proofs for Section 5

B.1 PrivCert-PF Algorithm

Algorithm 1 PrivCert-PF

Require: Query q, dataset R, proposal distribution Π_q, number of candidates m, threshold θ, margin τ, per-query privacy budget ε_0

Sample $s_1, \ldots, s_m \overset{\text{i.i.d.}}{\sim} \Pi_q$
for $j = 1, \ldots, m$ **do**
  Compute $\widehat{f}_{s_j}(R) = f_{s_j}(R) + Z_j$ using an ε_0-DP support query
  Set $L_{s_j} = \widehat{f}_{s_j}(R) - \tau$
  if $L_{s_j} \geq \theta$ **then**
    Emit s_j with certificate L_{s_j}
  else
    Abstain
  end if
end for
Optionally render emitted statements as post-processing

B.2 Privacy and Honesty

Privacy accounting. The proposal distribution Π_q is independent of R, so candidate sampling incurs no privacy cost. Conditioned on the sampled candidates, if support query j is ε_j-DP, sequential composition gives $\left(\sum_{j=1}^{m} \varepsilon_j\right)$-DP. The certificates, emit-or-abstain decisions, and any final rendering are post-processing and incur no additional privacy loss.

Proof of Theorem 3. For an evaluated statement s,

$$L_s = f_s(R) + Z - \tau.$$

Hence

$$L_s > f_s(R) \quad \Longrightarrow \quad Z > \tau,$$

so

$$\Pr[L_s > f_s(R)] \leq \Pr[Z \geq \tau] \leq \beta.$$

If $f_s(R) < \theta$ and s is emitted, then

$$f_s(R) + Z - \tau \geq \theta \quad \Longrightarrow \quad Z \geq \tau + \theta - f_s(R) > \tau,$$

and therefore

$$\Pr[\text{Emit}_s] \leq \Pr[Z \geq \tau] \leq \beta.$$

This proves certificate validity and per-candidate one-sided honesty.

Report-level control. Let B_j denote the event that candidate j is unsupported and emitted. Conditioned on the sampled candidates, Theorem 3 gives $\Pr[B_j] \leq \beta_j$. Thus

$$\Pr\left[\bigcup_{j=1}^{m} B_j\right] \leq \sum_{j=1}^{m} \beta_j.$$

In particular, uniform $\beta_j = \beta$ gives a report-level bound $m\beta$. No independence between candidate-level emission events is required.

B.3 Proof of Theorem 4

By linearity of expectation and identical sampling from Π_q,

$$\text{EmitRate} = \Pr_{s \sim \Pi_q, Z}\big[f_s(R) + Z - \tau \geq \theta\big].$$

For the upper bound, if $f_s(R) < \theta$, Theorem 3 gives emission probability at most β; otherwise we use the trivial bound 1. Hence

$$\begin{aligned}\mathrm{EmitRate} &\le p_\theta(R) + \beta\big(1 - p_\theta(R)\big) \\ &\le p_\theta(R) + \beta.\end{aligned}$$

For the lower bound, if

$$f_s(R) \ge \theta + 2\tau,$$

then

$$\theta + \tau - f_s(R) \le -\tau,$$

so

$$\Pr[s \text{ is emitted} \mid s] = \Pr[Z \ge \theta + \tau - f_s(R)] \ge \Pr[Z \ge -\tau] \ge 1 - \beta.$$

Averaging over $s \sim \Pi_q$ gives

$$\mathrm{EmitRate} \ge (1 - \beta) p_{\theta+2\tau}(R).$$

Combining the bounds proves

$$(1 - \beta) p_{\theta+2\tau}(R) \le \mathrm{EmitRate} \le p_\theta(R) + \beta(1 - p_\theta(R)) \le p_\theta(R) + \beta.$$

B.4 Proof of Corollary 1

Let

$$A := \frac{1}{\varepsilon_s} \log \frac{1}{2\beta}, \qquad k^\star := \lceil A \rceil.$$

Theorem 2 implies that emission probability $1/2$ requires distance at least $k^\star$.

For Laplace PrivCert-PF, define

$$\rho := \lceil n\theta \rceil - n\theta \in [0, 1).$$

At distance d from the largest unsupported support count,

$$f_s(R) - \theta = \frac{d - 1 + \rho}{n}.$$

With

$$\tau = \frac{1}{n\varepsilon_s} \log \frac{1}{2\beta} = \frac{A}{n},$$

the emission probability is at least $1/2$ iff $f_s(R) - \theta \ge \tau$, equivalently

$$d \ge A + 1 - \rho.$$

Thus the first such distance is

$$d_{\mathrm{PF}} = \lceil A + 1 - \rho \rceil \in \{k^\star, k^\star + 1\}.$$

The corresponding support count is $\lceil n\theta + A \rceil$, which is at most n by assumption; hence this distance is attainable.

B.5 Proof of Theorem 5

We prove the finite-sample bound

$$nw \ge \frac{\kappa m \log 2 + \log(1 - 8\beta)}{\varepsilon_{\mathrm{tot}}} - 2, \qquad \kappa := 1 - H_2(1/4),$$

where H_2 is binary entropy in bits. For fixed $0 < \beta < 1/8$ and any regime with $\varepsilon_{\mathrm{tot}} = o(m)$, this implies

$$w = \Omega\left(\frac{m}{n\varepsilon_{\mathrm{tot}}}\right).$$

Hard-instance construction. Let

$$t := \lceil n\theta \rceil - 1, \qquad h := \lceil n(\theta + w) \rceil, \qquad d := h - t.$$

Then

$$\frac{t}{n} < \theta, \qquad \frac{h}{n} \geq \theta + w,$$

and the feasibility assumption ensures $h \leq n$.

By the richness assumption, for every $v \in \{0,1\}^m$ there is a record r_v whose support pattern is v. Let $r_{\mathbf{0}}$ and $r_{\mathbf{1}}$ denote the all-zero and all-one patterns. Construct $R^{(0)}$ from t copies of $r_{\mathbf{1}}$ and $n - t$ copies of $r_{\mathbf{0}}$. For each v, obtain $R^{(v)}$ by replacing d copies of $r_{\mathbf{0}}$ by r_v. Then

$$d(R^{(0)}, R^{(v)}) \leq d,$$

and, for every j,

$$v_j = 0 \Longrightarrow f_{s_j}(R^{(v)}) = \frac{t}{n} < \theta, \qquad v_j = 1 \Longrightarrow f_{s_j}(R^{(v)}) = \frac{h}{n} \geq \theta + w.$$

Packing and decoding. By the Gilbert packing bound, there exists $\mathcal{C} \subseteq \{0,1\}^m$ with pairwise Hamming distance greater than $m/4$ and

$$|\mathcal{C}| \geq \frac{2^m}{\sum_{i=0}^{\lfloor m/4 \rfloor} \binom{m}{i}} \geq 2^{\kappa m}.$$

Let

$$X_j := \mathbf{1}\{j \in O\}, \qquad X = (X_1, \ldots, X_m).$$

For $v \in \mathcal{C}$, fine-grained certification gives

$$\Pr[X_j \neq v_j \mid R^{(v)}] \leq \beta$$

for every coordinate j, and therefore

$$\mathbb{E}[d_H(X, v) \mid R^{(v)}] \leq m\beta.$$

By Markov's inequality,

$$\Pr\Big[d_H(X, v) < \frac{m}{8} \,\Big|\, R^{(v)}\Big] \geq 1 - 8\beta.$$

Define

$$E_v := \Big\{x \in \{0,1\}^m : d_H(x, v) < \frac{m}{8}\Big\}.$$

Because codewords in $\mathcal{C}$ are more than $m/4$ apart, the events $\{E_v : v \in \mathcal{C}\}$ are pairwise disjoint.

Applying differential privacy. Group privacy and $d(R^{(0)}, R^{(v)}) \leq d$ imply

$$\Pr_{R^{(0)}}[X \in E_v] \geq e^{-d\varepsilon_{\mathrm{tot}}} \Pr_{R^{(v)}}[X \in E_v] \geq (1 - 8\beta)e^{-d\varepsilon_{\mathrm{tot}}}.$$

Summing over the disjoint decoding events,

$$1 \geq \sum_{v \in \mathcal{C}} \Pr_{R^{(0)}}[X \in E_v] \geq |\mathcal{C}|(1 - 8\beta)e^{-d\varepsilon_{\mathrm{tot}}},$$

hence

$$d\varepsilon_{\mathrm{tot}} \geq \log|\mathcal{C}| + \log(1 - 8\beta) \geq \kappa m \log 2 + \log(1 - 8\beta).$$

Finally,

$$d = \lceil n(\theta + w) \rceil - \lceil n\theta \rceil + 1 \leq \lceil nw \rceil + 1 \leq nw + 2.$$

Therefore

$$nw \geq \frac{\kappa m \log 2 + \log(1 - 8\beta)}{\varepsilon_{\mathrm{tot}}} - 2,$$

which proves Theorem 5.

C Additional Experimental Details and Diagnostics

This appendix provides implementation details and additional diagnostics for Section 6. We first describe the evaluation protocol, support model, proposal mechanisms, and privacy accounting, and then report additional comparisons that clarify the scope of the PrivCert guarantees.

C.1 Experimental Protocol and Metrics

Metrics. For each proposal-and-filter repetition k, we compute

$$\widehat{\text{UnsupportedEmit}}_k = \frac{\sum_j \mathbf{1}\{s_{kj} \text{ emitted}, f_{s_{kj}}(R) < \theta\}}{\sum_j \mathbf{1}\{f_{s_{kj}}(R) < \theta\}},$$

and report the mean across the S independent repetitions. This is the empirical analogue of the per-candidate quantity controlled by Theorem 3.

We also report the emitted fraction

$$\widehat{\text{EmitRate}}_k = \frac{1}{m} \sum_j \mathbf{1}\{s_{kj} \text{ emitted}\},$$

averaged across repetitions, together with

$$\widehat{\text{FalseEmission}} = \frac{\sum_j \mathbf{1}\{s_j \text{ emitted}, f_{s_j}(R) < \theta\}}{\sum_j \mathbf{1}\{s_j \text{ emitted}\}}, \qquad \widehat{\text{MeanSupport}} = \frac{\sum_j \mathbf{1}\{s_j \text{ emitted}\} f_{s_j}(R)}{\sum_j \mathbf{1}\{s_j \text{ emitted}\}}.$$

FalseEmission and MeanSupport are descriptive diagnostics rather than quantities directly controlled by Theorem 3. For free-text baselines, they are computed over parsed released claims.

Synthetic setup. The synthetic experiment isolates the private filter from semantic-matcher and proposal effects. We construct a finite statement universe $\mathcal{S} = \{s_1, \ldots, s_M\}$, assign each statement an oracle support value, and sample candidates uniformly from $\mathcal{S}$. Support values are drawn from a two-component mixture containing lower- and higher-support statements.

We sweep $\varepsilon_{\text{tot}} \in \{1, 4, 10, 20\}$, $\theta \in \{0.05, 0.1, 0.2\}$, $m \in \{50, 200, 1000\}$, $\beta \in \{0.01, 0.05, 0.1\}$, with five seeds per configuration.

Repetitions and hardware. Realistic proposal-and-filter experiments use five seeds per dataset–proposer pair and $m = 200$ candidates per seed. Most free-text baselines use three seeds; the WildChat/InvisibleInk Qwen configuration uses five. Experiments were run on Microsoft Azure VMs with NVIDIA H100 GPUs. Proposal-and-filter logs record the candidate, true operational support, noise draw, certificate, and emission decision, which are sufficient to reconstruct the reported metrics and theorem checks.

C.2 Support Evaluation, Proposals, and Privacy Accounting

Free-text claim extraction. Free-text baselines do not provide a fixed candidate pool or explicit emit-or-abstain decisions. We therefore extract atomic asserted statements from each released report using a fixed claim parser and score the resulting claims with the same support function used for PrivCert-PF. Claim extraction operates only on already released text and is therefore post-processing of the DP mechanism.

Operational support model. Our default support predicate is a frozen public NLI cross-encoder, `cross-encoder/nli-deberta-v3-base` (He et al., 2021). For a record r and statement s, r is used as the premise and s as the hypothesis. Inputs are tokenized with the native tokenizer and truncated to a maximum joint length of 384 wordpiece tokens.

Let $\ell_c(r, s)$ be the NLI logit for class c. We define

$$\sigma(r, s) = \mathbf{1}\left[\arg\max_c \ell_c(r, s) = \text{entailment}\right], \qquad f_s(R) = \frac{1}{|R|} \sum_{r \in R} \sigma(r, s).$$

The matcher is fixed independently of the private corpus. Because replacing one record changes at most one binary summand, $\Delta_{f_s} \le \frac{1}{|R|}$ under replace-one adjacency. This sensitivity statement

Table 4: **Privacy accounting for the main-table methods.** Pure-DP mechanisms also satisfy the displayed approximate-DP target.

Method	DP type	Composition unit	Sensitivity	Final guarantee
PRIVCERT-PF	pure DP	m support queries	$\Delta_f \le 1/\lvert R\rvert$	$\varepsilon_{\mathrm{tot}}$-DP
Document-level DP	pure DP	d_{doc} aggregates	Δ_{doc}	$\varepsilon_{\mathrm{tot}}$-DP
INVISIBLEINK	zCDP	private decoding	DClip sensitivity	$(\varepsilon_{\mathrm{tot}}, 10^{-5})$-DP
Unfiltered LLM	none	no private-data access	n/a	reference only

is independent of the semantic accuracy of the matcher; all guarantees are relative to the resulting operational support function.

Proposal mechanisms. The tag-template proposer samples structured values from a public schema over dimensions such as topic, aspect, and sentiment, and renders them through fixed templates. The public-LLM proposer instead receives only a public task description and generates candidate dataset-level claims. Both proposal mechanisms are fixed independently of R, so they incur no privacy cost. The proposer affects discovery and coverage but not the per-candidate honesty guarantee of the subsequent private verifier.

Privacy accounting. All privacy-constrained methods in the realistic experiments target $(\varepsilon_{\mathrm{tot}}, \delta) = (10, 10^{-5})$. Pure-DP mechanisms have $\delta = 0$ and therefore also satisfy this target.

For proposal-and-filter,

$$\varepsilon_0 = \frac{\varepsilon_{\mathrm{tot}}}{m}, \qquad Z_j \sim \mathrm{Lap}(b), \qquad b = \frac{m\Delta_f}{\varepsilon_{\mathrm{tot}}}.$$

The main experiments use the conservative margin

$$\tau = b \log \frac{1}{\beta},$$

which makes both required one-sided Laplace tail probabilities at most β.

The document-level baseline privately releases a fixed set of aggregate statistics and renders them with a public LLM. INVISIBLEINK uses its published private-decoding accountant; the zCDP budget is calibrated to the common $(10, 10^{-5})$-DP target, with parallel composition across disjoint reference batches.

C.3 ALTERNATIVE PRIVCERT MECHANISMS

Section 6.3 compares PRIVCERT-PF with alternative evidence constructions while holding the candidate space, support semantics, threshold, and honesty target fixed.

Contribution-bounded histogram. For candidates $s_1, \ldots, s_m$, define

$$\mathbf{f}(R) = \frac{1}{n} \sum_{r \in R} \big(\sigma(r, s_1), \ldots, \sigma(r, s_m)\big).$$

Its ℓ_1-sensitivity is at most m/n. If each record is instead allowed to contribute to at most c positive coordinates, the capped vector satisfies

$$\Delta_1(\widetilde{\mathbf{f}}_c) \le \frac{\min(m, 2c)}{n},$$

and therefore uses Laplace scale

$$b_c = \frac{\min(m, 2c)}{n\varepsilon_{\mathrm{tot}}}.$$

Contribution bounding can reduce noise but may remove genuine support. Because only positive contributions are removed, $\widetilde{f}_{c,s}(R) \le f_s(R)$, so the resulting lower certificates remain conservative with respect to the original support function. The main comparison uses $c = 16$.

Sparse-vector select-then-certify. The SVT variant allocates $\varepsilon_{\mathrm{tot}} = \varepsilon_{\mathrm{select}} + \varepsilon_{\mathrm{cert}}$, uses the first budget to select at most k candidates, and certifies each selected candidate independently with budget

Table 5: **Proposer comparison on realistic corpora.** Only the proposal distribution changes; the private verifier is fixed.

Dataset	Proposer	#out	EmitRate	UnsupEmit	FalseEmit	MeanSupport
TAB	public LLM	3	0.003	0.003	1.000	0.006
	tag-template	404	0.404	0.008	0.010	0.545
WildChat	public LLM	24	0.024	0.001	0.042	0.132
	tag-template	220	0.220	0.004	0.014	0.288
Yelp	public LLM	18	0.018	0.002	0.111	0.124
	tag-template	749	0.749	0.000	0.000	0.239

$\varepsilon_{\text{cert}}/k$. The final certification stage therefore retains the one-sided honesty argument of Theorem 3. Theorem 4 does not apply unchanged because SVT introduces dependent selection, a shared noisy threshold, early stopping, and a cap on positive outputs. We sweep $k \in \{5, 10, 20, 40\}$, $\frac{\varepsilon_{\text{select}}}{\varepsilon_{\text{tot}}} \in \{0.25, 0.50, 0.75\}$.

Gaussian backend. To verify that the reporting interface is not specific to pure-DP Laplace noise, we also use

$$\widehat{f}_s(R) = f_s(R) + Z_s, \qquad Z_s \sim \mathcal{N}(0, \sigma_G^2),$$

with σ_G calibrated so that the complete multi-statement release satisfies $(\varepsilon_{\text{tot}}, \delta) = (10, 10^{-5})$-DP. The certificate margin is

$$\tau_G = \sigma_G \Phi^{-1}(1 - \beta),$$

so that

$$\Pr[Z_s \geq \tau_G] = \beta.$$

The same one-sided argument as in Theorem 3 therefore applies.

These alternatives demonstrate that the PRIVCERT contract is not specific to a particular DP primitive; their coverage differs because they exploit different sensitivity and selection structures.

C.4 PROPOSER COMPARISON

We hold the support model, private verifier, privacy budget, and number of candidates fixed and change only the proposal distribution. Both proposers use $(\varepsilon_{\text{tot}}, \delta) = (10, 10^{-5})$, $m = 200$, with five seeds per dataset.

UnsupportedEmit remains below 0.01 across all six cells, while coverage changes substantially. This illustrates the distinction between discovery and certification: changing the proposer changes which certifiable statements are found without changing the verifier's one-sided honesty property.

C.5 DP SYNTHETIC DATA AS A BOUNDARY CASE

DP synthetic data provides a different reporting contract: a private synthetic corpus $\widetilde{R}$ is released first, and subsequent reporting is performed on that proxy rather than directly on R.

We evaluate an Aug-PE-style backend on Yelp using the same private split of 1800 records, public statement schema, semantic matcher, and reporting thresholds. The synthetic pipeline follows the main structure of Private Evolution (Xie et al., 2024), using private voting, Gaussian-noised selection, and LLM-based variation. The best of four evaluated configurations uses population size $N = 1000$.

The deployable synthetic-data rule emits a statement when $f_s(\widetilde{R}) \geq \theta$. For evaluation only, we additionally compute its original-corpus support $f_s(R)$; this quantity is never used by the mechanism. We report

$$\text{SupportedRecall} = \frac{|\{s : f_s(R) \geq \theta,\ s \text{ emitted}\}|}{|\{s : f_s(R) \geq \theta\}|}.$$

At $\theta = 0.10$, the synthetic backend achieves similar coverage but substantially larger original-data UnsupportedEmit. The repetition units differ—a synthetic-data seed regenerates the complete proxy,

Table 6: **Matched Yelp comparison with a DP synthetic-data backend.** UnsupportedEmit for synthetic data is evaluated against support in the original private corpus.

θ	Method	EmitRate	UnsupEmit	Runs w/ any unsup.	SupportedRecall
0.05	PRIVCERT-PF	0.823	0.000	$0/150$	0.979
0.05	Aug-PE-style ($N = 1000$)	0.800	0.000	$0/5$	0.952
0.10	PRIVCERT-PF	0.695	0.001	$2/150$	0.938
0.10	Aug-PE-style ($N = 1000$)	0.676	0.092	$4/5$	0.881

Table 7: **Robustness of PRIVCERT-PF to a Qwen3-8B support model.** Qwen3-8B replaces the default NLI matcher as the operational support predicate. "In interval" reports finite-run consistency with the expectation interval of Theorem 4.

Dataset	UnsupportedEmit	NLI–Qwen Jaccard	In interval
TAB	0.0033	0.94	5/5
WildChat	0.0023	0.33	5/5
Yelp	0.0037	0.78	5/5

whereas a proposal-and-filter repetition resamples final DP noise—so the run-level counts should not be interpreted as matched failure-probability estimates.

The distinction is semantic. Support in $\widetilde{R}$ certifies the proxy, not automatically the original dataset. An original-data lower certificate would require an additional fidelity guarantee such as

$$\Pr\left[\sup_{s\in\mathcal{S}}\big(f_s(\widetilde{R}) - f_s(R)\big) \le \gamma_+\right] \ge 1-\eta.$$

Under this condition, $f_s(\widetilde{R}) - \gamma_+ \ge \theta$ would imply $f_s(R) \ge \theta$ except with probability at most η. An oracle post-hoc correction based on original-corpus supports removes unsupported emissions in this experiment but reduces EmitRate to approximately 0.14–0.28; because it inspects the private-data oracle, it is not a deployable mechanism.

C.6 ROBUSTNESS TO THE SUPPORT MODEL

The guarantees of PRIVCERT are defined relative to a fixed support function f_s. We therefore separate two questions: whether the private certification mechanism remains well calibrated when the support model changes, and whether the identities of supported statements are stable across support models.

Alternative NLI matchers. We repeat proposal-and-filter with three frozen public NLI matchers: the default DeBERTa cross-encoder, a second DeBERTa model trained on MNLI/FEVER/ANLI, and BART-large-MNLI. All are fixed independently of the private corpus.

Observed UnsupportedEmit remains below the target β for every matcher, consistent with Theorem 3. The emitted content can nevertheless differ substantially. On Yelp, 92.7% of the union of statements emitted by any matcher is shared by all three; on WildChat, only 19.3% is shared by all three and 39.0% is emitted by only one. Thus mechanism-level honesty is distinct from semantic agreement between support models.

Qwen3-8B as the support model. We additionally replace the NLI predicate with a frozen Qwen3-8B judge. The model and prompt are fixed independently of the private corpus, thinking is disabled, and each record–statement pair is deterministically mapped to a binary support decision. The resulting empirical-frequency support retains replace-one sensitivity at most $1/|R|$.

Importantly, we rerun PRIVCERT-PF using the Qwen-defined support function rather than merely re-scoring outputs produced under the NLI support model.

UnsupportedEmit remains well below $\beta = 0.05$, and all $15/15$ finite-run EmitRates are empirically consistent with the interval predicted for the expectation by Theorem 4. The emitted sets nevertheless differ, especially on WildChat. Thus the mechanism-level behavior persists under an LLM-based support model, while coverage and certified content remain support-model dependent.

Table 8: **Cross-evaluation of exactly recoverable frozen Table 1 outputs.** Released claims are fixed and only the operational support evaluator changes from NLI to Qwen3-8B. Qwen FalseEmission is a descriptive diagnostic and carries no β-guarantee.

Dataset	Method	NLI FalseEmit	Qwen FalseEmit
TAB	Document-level DP	0.894	0.470
	PRIVCERT-PF	0.010	0.040
WildChat	Document-level DP	0.729	0.857
	INVISIBLEINK (Qwen)	0.619	0.929
	PRIVCERT-PF	0.014	0.241
Yelp	Document-level DP	0.803	0.239
	PRIVCERT-PF	0.000	0.144

Table 9: **Coverage cost of whole-report control.** The family-wise setting uses $\beta_j = \beta_{\text{tot}}/m$ with $\beta_{\text{tot}} = 0.05$.

Dataset	Per-candidate EmitRate	Family-wise EmitRate
TAB	0.404	0.357
Yelp	0.749	0.565
WildChat	0.220	0.169

Cross-evaluation of frozen free-text outputs. The previous experiment changes the support function and reruns the certification mechanism. We separately test whether the high FalseEmission of free-text baselines in Table 1 is specific to the default NLI evaluator.

For this diagnostic, released claims are held fixed and only the evaluator is changed from NLI to the same frozen Qwen3-8B judge. Seven of the fifteen dataset–method cells could be recovered exactly. Each recovered cell first reproduces its original Table 1 NLI FalseEmission before Qwen scoring. No model outputs were regenerated to fill unavailable cells.

Among the exactly recoverable cells, the ordering between PRIVCERT-PF and the available free-text baselines is preserved on all three datasets. The gap remains large on TAB and WildChat but narrows substantially on Yelp. Thus the qualitative evidence-gap diagnostic is not explained solely by the default NLI matcher, while its magnitude is clearly support-model dependent.

The Qwen scores assigned to the frozen PRIVCERT-PF outputs do not carry the original β-honesty guarantee because those outputs were certified under the NLI-defined support function. This cross-evaluation should therefore be interpreted only as a robustness diagnostic. The remaining Table 1 cells are omitted because their exact frozen claims were unavailable.

C.7 WHOLE-REPORT GUARANTEE

The main experiments use the per-candidate guarantee of Theorem 3. If instead the complete report must contain no unsupported statement except with probability at most β_{tot}, the union bound in Appendix B allows failure targets satisfying $\sum_{j=1}^{m} \beta_j \le \beta_{\text{tot}}$. With uniform allocation, $\beta_j = \frac{\beta_{\text{tot}}}{m}$. For the conservative Laplace calibration used in the main experiments, the certificate margin becomes

$$\tau_{\text{FW}} = \frac{m\Delta_f}{\varepsilon_{\text{tot}}} \log \frac{m}{\beta_{\text{tot}}},$$

which is larger than the per-candidate margin and therefore reduces coverage.

The stronger report-level guarantee reduces coverage on all three datasets. No unsupported report is observed in these evaluated runs, but the formal statement remains

$$\Pr[\text{any unsupported emission}] \le 0.05.$$

Thus per-candidate calibration provides the higher-coverage operating point used in the main evaluation, while family-wise calibration is available when the guarantee must apply to the complete released report.