UTILITY, SENSITIVE-CONTENT EXCLUSION, AND PRIVACY IN GRAPH-BASED SUMMARISATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Summarizing sensitive documents requires useful content, control over sensitive source spans, and a precise privacy guarantee. We develop a framework that separates these requirements and connects each to testable evidence in a source-backed graph summariser. We prove selection privacy under fixed public support and bounded score sensitivity, and construct counterexamples showing why stable filtering alone cannot extend this guarantee to private extractive outputs. We establish source-level exclusion under detector coverage and provenance conditions, and derive a tight objective-headroom bound on selection gains over uniform feasible summaries. Across eight corpora and 355,200 selection records, all 135 Holm-significant exponential-versus-frontier effects are positive. Controlled disease-label interventions isolate the effect of protection; evaluations on 100 biomedical abstracts and 48 clinical cases quantify the gap between excluding detected spans and covering every sensitive span. Larger protection radii reduce observed exposure at the cost of more empty releases. A 240-cell provider-adapted generation study complements these results with coverage gains and explicit bounds for its 11 failed cells. Together, the analysis and experiments identify which evidence supports each release claim and which design choices separate content protection from differential privacy.

1 INTRODUCTION

Summaries make document collections easier to use, but releasing a summary also selects which information becomes visible. In sensitive settings, three requirements therefore arise: retain useful content, exclude protected source content, and control what the release reveals about a contributor. These requirements concern different objects: an objective value, source spans, and output distributions on neighbouring inputs. Treating them as interchangeable can produce a convincing evaluation of the wrong guarantee.

We develop a unified account of these requirements through a source-backed sentence-graph summariser. The system filters protected regions, selects connected sentences, and renders them with explicit provenance. This structure lets us relate mathematical assumptions to individual pipeline operations and measurable outcomes. Our central insight is that the evidence must follow the claim: objective headroom governs selection gains, detector coverage and provenance govern source exclusion, and output support and the complete observation process govern differential privacy.

We show that stable survivor labels can coexist with loss of private output support, whereas complete detector coverage supports source exclusion even with false positives. Our experiments quantify selection gains and suppression costs. Together, these results connect component guarantees to pipeline behavior and identify the architectural choices governing release, a distinction that matters for learning systems combining retrieval, filtering, and generation.

Contributions.

- **A framework for distinct release guarantees.** We connect utility, sensitive-content exclusion, and differential privacy to explicit assumptions and evidence in graph-based summarization, including a tight headroom bound for gains over uniform feasible selection.

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.

- **A characterization of guarantee boundaries.** We prove conditional selection privacy and source-exclusion results, and construct counterexamples separating survivor stability from common output support and initial headroom from later evolution gains.
- **An empirical account of selection and protection.** Eight-corpus experiments with 355,200 records, controlled label interventions, and independently annotated medical evaluations quantify selection gains, coverage gaps, and suppression costs. A provider-adapted generation study measures complementary distributional coverage.

2 Related Work

Private selection and data-dependent support. The exponential mechanism selects from a fixed output universe using bounded-sensitivity scores McSherry and Talwar (2007). Restricted sensitivity extends queries from a restricted input family to obtain global sensitivity guarantees Blocki et al. (2013). We use these established principles to distinguish score stability from support stability in extractive summarization. Our counterexamples identify the specific step at which contributor deletion defeats a score-only privacy argument. Permute-and-flip improves expected score for a fixed score vector McKenna and Sheldon (2020); its equivalence to report-noisy-max uses exponential noise Ding et al. (2021). We evaluate per-round alternatives within adaptive selection, where single-round comparisons alone leave trajectory ordering open.

Summarization and generation. Submodular objectives provide an established approach to document summarization Lin and Bilmes (2011). We use facility location to connect a simple headroom identity to feasible-family comparisons and sampling uncertainty. Augmented Private Evolution (Aug-PE) generates synthetic text through model APIs Xie et al. (2024). Our Copilot-backed adaptation measures distributional coverage alongside source-backed extraction, making the different evaluation endpoints explicit. An illustrative LexRank control using Sumy Sumy contributors (n.d.) appears in Appendix C.

Implementation-aware privacy. Finite-precision analysis Mironov (2012), exact base-2 exponential sampling Ilvento (2020), and privacy counterexample testing Ding et al. (2018) establish the importance of implementation and observation models. Our public-label construction and transcript analysis apply this perspective to a concrete release boundary (Appendix A). The contribution is the connection between these requirements and graph-based summarization, supported by proofs, counterexamples, and matched empirical checks.

3 Setting and Threat Model

Input is a document string, optional entity spans and contributor attributions, and configuration parameters. Sentence splitting produces vertices with source intervals, embeddings and clipped weights $I(v) \in [0, 1]$. Similarity edges join pairs exceeding τ; consecutive sentences receive discourse edges. An individual contributes at most c retained vertices. Unknown ownership is represented by a public placeholder; person-level privacy statements require trusted attribution.

Let T be the true sensitive source spans and D the detected spans. Our content policy includes direct identifiers, disease names, diagnoses and medical conditions. Generic medical discussion is included conservatively; this definition is broader than legal identifiability. Other quasi-identifiers are optionally excluded. We evaluate source exclusion against T and run the filter against D, keeping annotation truth separate from observed labels.

An adversary may observe the summary and know other contributions. Contributor neighbours remove all vertices attributed to one individual, retaining labels and weights of survivors. This graph-level adjacency fixes surviving quantities; rebuilding them from raw text is a different operation. Differential privacy requires $\Pr[M(X) \in B] \leq e^{\varepsilon} \Pr[M(X') \in B] + \delta$ for every measurable output event B and neighbouring pair, with $\delta = 0$ for pure DP. Debug graphs, detected spans, scores, source offsets, and realized budget traces remain inside the trusted boundary.

4 METHOD

The graph pipeline separates protection, selection, and realization so that each guarantee can be traced to the operation that supplies it.

Filter. Mark each vertex whose provenance overlaps D. Exclude these seeds and their radius-r graph neighbourhood. The remaining vertices form C. Explicit disease/condition labels map to the protected tier independently of the quasi-identifier toggle. Detection supplies the spans; the policy determines their protection.

Select. Maximize facility-location coverage

$$f(S) = \sum_{v \in V} I(v) \max_{s \in S} \mathrm{sim}(v, s), \qquad f(\emptyset) = 0.$$

At each of at most k rounds, sample a candidate with weight $\exp(\varepsilon_t g_t(v)/(2\Delta u))$, where g_t is its marginal gain. The connected variant chooses from the current frontier after its first round. Basic accounting allocates $\varepsilon_t = \varepsilon/k$; an empty frontier stops selection. Throughout the private-text experiments, ε denotes the nominal selection parameter. Section 5 identifies the additional conditions under which it is a privacy budget.

Realize and verify. Sort selected sentences by source position, render with fixed templates, and check source provenance and protected-region membership. Invalid output is suppressed. Source offsets and verification certificates record the origin and protection status of the selected sentences inside the trusted boundary. Learned medical detectors use checked source offsets and the same protection policy, including negated, historical, and generic disease mentions (Appendices B and D).

5 PRIVACY AND SOURCE-EXCLUSION GUARANTEES

We establish two distinct guarantees: distributional privacy for selection over public outputs, and sensitive-span exclusion for source-backed extraction. The counterexamples show why preserving one guarantee's premises need not preserve the other's.

5.1 SELECTION PRIVACY AND THE OUTPUT-SUPPORT BOUNDARY

The following application of exponential-mechanism composition states sufficient conditions for privacy of the entire selection transcript. Its essential requirement is a common public output universe at every shared history. Randomness is fresh and its seed remains undisclosed.

Theorem 1 (Public-universe selection)**.** *Let $X \sim X'$ differ in one private contributor. Candidate contents and the candidate graph are public and fixed. For each common released history h, the feasible candidates $A(h)$ and stop rule are identical on both inputs, and $|g_X(h,a) - g_{X'}(h,a)| \le \Delta u$ for all $a \in A(h)$. At each nonterminal history $A(h)$ is finite and nonempty, $0 < \Delta u < \infty$, and the positive round budget is a public function of that history. Sequential exponential choices with $\sum_t \varepsilon_t \le \varepsilon$ on every transcript are ε-DP. Applying a fixed rendering/verification kernel independent of X preserves the guarantee.*

Proof. With $\alpha_t = \varepsilon_t/(2\Delta u)$, the ratio of corresponding exponential weights is at most $e^{\varepsilon_t/2}$. Summing over the common $A(h)$ bounds the inverse ratio of partition functions by the same factor. Each conditional output ratio is at most e^{ε_t}. Multiplication along a transcript gives at most $e^{\sum_t \varepsilon_t}$. Summation over transcripts leading to an output, or integration against a fixed post-processing kernel, proves the result. □

For fixed public candidates and similarities, and private voter contributions of at most c terms with weights bounded by $I_{\max}$, a facility-location marginal gain changes by at most $cI_{\max}$: each removed voter's incremental maximum lies in $[0, I_{\max}]$. This supplies the theorem's score bound when survivor weights and similarities remain fixed. Recomputing them requires a separate sensitivity bound. Score sensitivity leaves a separate question: can the same output still occur after a contributor is removed?

Proposition 2 (Survivor stability is not common support)**.** *At $r = 0$, inherited sensitivity labels yield $C_0(G - c) = C_0(G) \cap V(G - c)$. This survivor-stability property is insufficient for pure DP under contributor deletion. For every $r \geq 1$, survivor stability can also fail.*

Proof. At radius zero a surviving vertex's seed status is unchanged. For the privacy counterexample, take one eligible private sentence and one eligible public sentence with equal singleton scores. At $k = 1$ the private sentence is output with probability $1/2$ before deletion and zero after. No finite ε satisfies the DP inequality for that event. For $r \geq 1$, connect a private PII seed to a public vertex by a path of length at most r. Deleting the seed can make that vertex eligible. □

Proposition 2 separates two requirements that a filter-stability argument can conflate. Survivor stability preserves eligibility among remaining vertices; pure DP requires common output support. Copying a private sentence after selecting its ID retains the support dependence. For approximate DP, the probability of an exclusive output must be at most δ. Radius-two filtering has a further structural dependence: deleting a non-identifier bridge on a public-seed–bridge–public path can unprotect its final vertex, even with fixed identifier labels. These examples identify support and graph structure as distinct obligations beyond score sensitivity.

A public-output construction. Fix a public vocabulary and clip each person's vote for each string to $[0, 1]$. The summed votes have sensitivity one, so exponential selection of a public string satisfies Theorem 1 in exact arithmetic. This construction changes the release to public vocabulary content rather than private sentences. Appendix A gives an exact rational implementation and characterizes privacy of a snapshot reader's complete transcript under explicit publication and environment assumptions. Appendix K analyzes a removal-only stability diagnostic separately.

5.2 Sensitive-content exclusion

For source-backed extraction, provenance gives a direct content guarantee. The next theorem separates exclusion of observed sensitive spans from exclusion of all true sensitive spans; detector coverage is the condition connecting them.

Theorem 3 (Label-relative and truth-level exclusion)**.** *Suppose output content is confined to the provenance spans of selected vertices, and selection and verification exclude every vertex in the protected region seeded by D. No released source span overlaps D. If every vertex whose provenance overlaps T is also seeded by D, no released source span overlaps T.*

Proof. Any vertex whose provenance overlaps D is a seed, hence protected at every nonnegative radius and unavailable for release. Under the second premise this includes every vertex overlapping T. Confinement of rendered content to the selected provenance completes both statements. □

The theorem permits false positives: they cost eligible content but preserve exclusion. Complete recall on release-relevant spans is sufficient for truth-level protection. The guarantee concerns overlap with source spans; semantic inference and reidentification are different endpoints. A probabilistic coverage premise yields a corresponding exposure bound.

Proposition 4 (Document-level detector error)**.** *If the coverage or provenance premise of Theorem 3 fails with probability at most η, then true-sensitive source overlap in the release has probability at most η.*

Proof. On the complement of the premise-failure event, Theorem 3 forbids overlap. Thus the overlap event is a subset of that failure event. □

Proposition 4 makes complete-document coverage the relevant evidence for an exposure bound, allowing arbitrary dependence among token errors. Its η bounds source-overlap probability, not the distributional divergence in DP. Finite-sample token recall alone cannot supply a uniform document-level η; our medical evaluations therefore report both mention-level accuracy and complete-document coverage.

6 OBJECTIVE HEADROOM

How much can a better selector improve the objective on a fixed feasible family? We answer this through the gap between uniform selection and the optimum. This gives a reference scale for interpreting small empirical gains. For finite nonempty feasible family $\mathcal{F}$, nonnegative f and $f^\star = \max_{S\in\mathcal{F}} f(S) > 0$, define $\rho = 1 - \mathbb{E}[f(U)]/f^\star$, where U is uniform on $\mathcal{F}$. If the family is empty or its maximum is zero, the normalized quantity is undefined.

Proposition 5 (Tight headroom bound)**.** *Every $\mathcal{F}$-valued rule A satisfies $\big(\mathbb{E}[f(A)] - \mathbb{E}[f(U)]\big)/f^\star \le \rho$, with equality exactly when A returns a maximizer almost surely. If $\mathbb{E}[f(U)] > 0$, then*

$$\mathbb{E}[f(A)]/\mathbb{E}[f(U)] \le 1/(1-\rho).$$

Proof. Write $\mu = \mathbb{E}[f(U)] = (1-\rho)f^\star$. Since $f(A) \le f^\star$, subtracting μ and dividing by $f^\star$ proves the first inequality and its equality condition. Dividing $\mathbb{E}[f(A)] \le f^\star$ by positive μ proves the ratio bound. □

The bound is exact: ρ is the maximum normalized expected gain over uniform selection. Small headroom can coexist with a large score range, as in $\{0, 1, \ldots, 1\}$. Filtering can also increase or decrease headroom: from $\{0, 9, 10\}$, retaining $\{0, 10\}$ raises it, whereas retaining $\{9, 10\}$ lowers it. Thus the effect of filtering depends on the retained objective values, not just the candidate count.

The comparison must use the same feasible family. Frontier-random selection induces a different distribution from uniform connected-subset sampling, and early-stopped outputs differ from full-k sets. We distinguish these baselines in the experiments and count failed uniform rejection draws separately. Appendix E supplies the standard unconstrained greedy utility bound.

We also derive a sampling interval using a known score range and any valid feasible lower bound on the optimum (Theorem 8). It supports low-headroom decisions without an exact optimum. The fixed-family restriction matters for generation: a public mutation can turn zero initial headroom into a $2/3$ gap among possible child outputs (Appendix F.2). Initial saturation therefore supplies no general bound on later evolution gains without assumptions connecting parent and child utilities.

7 EXPERIMENTAL DESIGN

The evaluation asks three questions: how much utility selection gains over its baselines, how detector coverage determines source exposure, and how synthetic evolution changes distributional coverage. Each experiment uses the endpoint appropriate to its claim.

Corpora. We use synthetic single-topic and six-topic documents, BillSum, GovReport, the publisher-described synthetic and human-validated ai4privacy 300k corpus, and the OpenReview, Yelp and PubMed training collections distributed with Aug-PE Xie et al. (2024). Each contributes 40 documents; sampling excludes empty rows and exact trimmed-text duplicates. The five ordinary-text corpora lack independent PII annotations and supply utility measurements. Controlled interventions and the separately annotated NCBI Disease and E3C datasets supply the source-exposure endpoints. Dataset provenance is recorded in Appendix C.

Protocol. We retain the first 24 sentences, use hashing embeddings, fixed $\tau = 0.6$, discourse weight 0.2, contribution cap $c = 3$, and seeds $0, \ldots, 9$. The core matrix crosses $k \in \{2, 4\}$, $r \in \{0, 1, 2\}$ and $\varepsilon \in \{0.1, 1, 8, 64, 512\}$. Methods are exponential connected greedy, non-private connected greedy, frontier random and uniform feasible selection. Greedy runs once per cell. A fixed threshold holds the graph-construction rule constant across corpora.

Ablations. At $k = 2, r = 1, \varepsilon = 1$, vary connectivity, objective, composition, quasi-identifier exclusion, contribution cap, discourse edges, separator filter or per-round mechanism, one factor at a time. Appendix J reports the intervals, keeping redundancy-penalized utility separate from facility-location scores.

Medical intervention. After truncation append one fixed sentence containing *diabetes*, with known source span and contributor attribution. Compare protected disease labelling, a missed disease label, and the legacy quasi-identifier policy with that toggle off. All other labels and attribution remain

Table 1: Facility-location utility, $k = 2, r = 1, \varepsilon = 1$, 40 documents per corpus. Within each row, the best (highest) mean is bold and second-best underlined. Ties in displayed values share ranks; formatting does not denote significance.

Corpus	Exponential (Ours)	Greedy	Frontier random	Uniform feasible
Synthetic single	4.639	**5.136**	4.616	<u>4.761</u>
Synthetic multi	<u>4.808</u>	**5.629**	4.797	4.727
BillSum	<u>4.440</u>	**5.957**	4.099	4.386
GovReport	6.164	**7.364**	6.104	<u>6.181</u>
Annotated synthetic	1.758	<u>2.214</u>	1.735	**2.372**
OpenReview	4.792	**5.801**	4.714	<u>4.814</u>
Yelp	2.684	**3.040**	2.679	<u>2.689</u>
PubMed	<u>3.565</u>	**4.087**	3.514	3.511

fixed. At $k = 2, \varepsilon = 1$, this intervention isolates the effect of protecting the known span, with up to 25 sentences per document. The exposure endpoint is overlap with the injected span; natural-text detector accuracy is assessed separately.

Statistics. Average seeds within each document before aggregation. Report percentile bootstrap intervals over documents (10,000 resamples). Core exponential-minus-frontier differences use paired sign-flip tests and Holm adjustment across 240 comparisons. The sign-flip null assumes exchangeable signs. Empty releases retain zero utility; failed uniform draws remain missing and are counted. Raw records retain every scheduled cell. Medical-policy differences versus oracle use a separate Holm family of 48 comparisons, with document-level bootstrap intervals on the source-overlap rate difference.

8 Results and Analysis

8.1 What does selection gain?

The 355,200-record evaluation shows a budget-dependent utility advantage for exponential selection over frontier random: all 135 Holm-significant effects among 240 comparisons are positive. At the representative $k = 2, r = 1, \varepsilon = 1$ slice (Table 1), the advantage is small on several corpora, and non-private greedy has higher utility on every corpus. This locates the low-budget selector relative to both a random baseline and a score-maximizing greedy rule.

Figure 1 shows mean exponential utility normalized by mean non-private greedy utility. Increasing the nominal budget concentrates choices on high marginal gains, and the observed ratios approach the greedy reference. Uniform feasible selection and frontier random give different reference values because they sample different distributions. Uniform draws succeed in 86,950 of 96,000 scheduled attempts; Appendix G characterizes the unavailable full-k families. Per-comparison effects, intervals, p-values, and document counts accompany the raw results.

Headroom as a decision criterion. Exact finite-graph evaluation tests whether the sampling interval can identify nearly saturated objectives. Among nine defined held-out graph/domain cells, four have exact headroom at most 0.1. The fixed decision rule identifies two, abstains on the other two, and makes no false low-headroom decision. In the identified cells, the maximum normalized gain over uniform is at most 0.00154. The mean interval width is 0.231, showing why the rule also abstains. This diagnostic demonstrates selective identification of low-gain families on the synthetic instances; Appendix F gives the full design and Figure 2.

8.2 When do labels prevent source exposure?

Controlled protection. Correct disease labels exclude the injected sensitive span in every evaluated cell. Missed and legacy labels leave it eligible even when the label-relative verifier passes (Table 2). At radius zero, the missed-label overlap rate is 22.75% on the annotated synthetic sample and 2.25% on OpenReview. BillSum and GovReport have zero observed overlap because the eligible injected

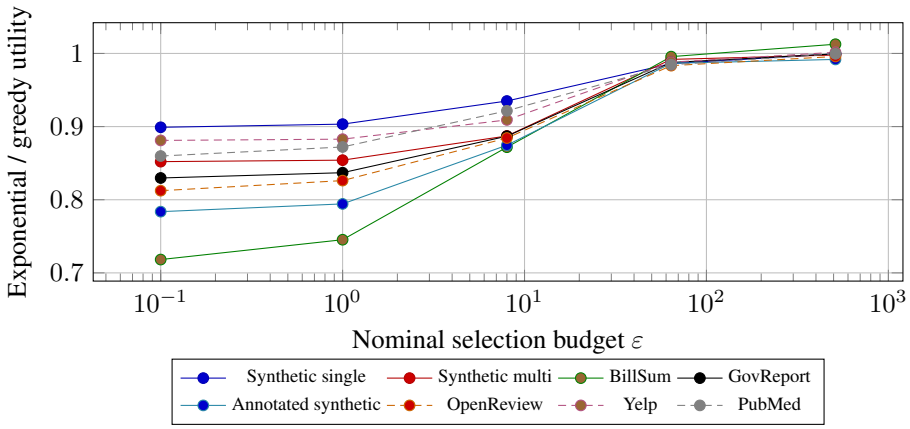

Figure 1: Mean exponential-to-greedy utility ratio at $k = 2, r = 1$ as the nominal selection parameter varies. Document-level intervals accompany the raw means in the supplied summary.

Table 2: Injected disease source-overlap rate (percent), $k = 2, r = 0, \varepsilon = 1$, 40 documents and 10 seeds. Oracle, missed and legacy vary observed label policy while holding truth fixed. Among the three overlap rates in each row, best (lowest) is bold and second-best underlined; ties share ranks. Candidate counts describe protection cost and are unranked.

Corpus	Oracle	Missed	Legacy	$\lvert C\rvert$ oracle	$\lvert C\rvert$ missed
Synthetic single	**0.00**	3.25	3.25	11.00	12.00
Synthetic multi	**0.00**	0.25	0.25	21.82	22.82
BillSum	**0.00**	**0.00**	**0.00**	23.80	24.80
GovReport	**0.00**	**0.00**	**0.00**	24.00	25.00
Annotated synthetic	**0.00**	22.75	22.75	7.88	8.72
OpenReview	**0.00**	2.25	2.25	19.27	20.27
Yelp	**0.00**	30.75	30.75	7.70	8.70
PubMed	**0.00**	16.50	16.50	10.60	11.60

sentence is not selected in these trials. Holding the source and other labels fixed isolates the role of coverage in Theorem 3.

Coverage on biomedical abstracts. Learned detection reduces source exposure while leaving substantial gaps in complete-document coverage. On 100 NCBI Disease test abstracts with 960 gold mentions Dogan et al. (2014), the pinned biomedical detector d4data (n.d.) has 15.4% exact-span precision, 10.0% recall, and complete gold-mention coverage on 0/100 documents (Table 3). Thresholds are fixed before test evaluation. Matched unconstrained selection at $k = 2, \varepsilon = 8, \tau = 0.6$ holds the seed and graph fixed across oracle, regex, and medical labels. At radius zero, medical labels reduce exposure from 65/100 to 50/100 abstracts (medical Wilson interval [40.4%,59.6%]); oracle labels exclude every gold span.

The cost of a larger protection region. At radii one and two, biomedical exposure falls to 39% and 28%, while empty releases rise to 22% and 44%. Radius expansion therefore trades retained context for fewer observed exposures, rather than improving the detector itself. Appendix B reports utility and safe-node suppression; Appendix I reports the separate injected-span radius interventions.

Coverage on clinical cases. A disease-specific BioBERT model evaluated on the 48 official E3C test cases Zanoli et al. (2023) achieves 46.9% exact clinical-segment precision, 64.0% recall, and complete coverage on 3/48 cases (95% Wilson interval [2.1%,16.8%]). E3C clinical entities are broader than disease labels; the evaluation retains nested and discontinuous segments. Medical source exposure is 8/4/2 cases at radii 0/1/2, compared with 33 under regex labels and zero under oracle labels (Appendix D).

The two evaluations show why exposure reduction and whole-document protection must be reported separately. Neither detector meets the frozen targets of 80% precision and a lower 95% document-

Table 3: NCBI Disease test set, $r = 0$. Precision/recall are percentages; other columns count documents out of 100. Complete requires coverage of every gold mention. Within each column, best is bold and second-best underlined (higher precision/recall/complete, lower exposed/empty); ties share ranks and missing values are unranked. All rows share the graph pipeline and differ in their oracle or detector labels.

Labels	Precision	Recall	Complete	Exposed	Empty
Oracle	**100.0**	**100.0**	**100**	**0**	3
Regex	–	0.0	<u>0</u>	65	**0**
Medical	<u>15.4</u>	<u>10.0</u>	<u>0</u>	<u>50</u>	<u>2</u>

Table 4: Evolved-minus-initial generation coverage on fixed cohorts. CC intervals concern successful runs. Scheduled bounds permit each missing effect to lie in $[-1, 1]$ and describe the finite scheduled mean, not a population confidence interval. Distinct corpora and uncertainty intervals are not ranked.

Corpus	Complete	Difference	CC 95% CI	Scheduled bounds
Single-topic	30/30	0.120	[0.102, 0.138]	[0.120, 0.120]
Multi-topic	30/30	0.093	[0.073, 0.114]	[0.093, 0.093]
BillSum	28/30	0.090	[0.073, 0.106]	[0.017, 0.150]
GovReport	29/30	0.084	[0.073, 0.095]	[0.048, 0.114]
ai4privacy	29/30	0.007	[0.002, 0.011]	[-0.027, 0.040]
OpenReview	26/30	0.040	[0.032, 0.048]	[-0.098, 0.168]
Yelp	29/30	0.032	[0.026, 0.038]	[-0.002, 0.064]
PubMed	28/30	0.087	[0.076, 0.099]	[0.015, 0.148]

coverage bound of 95%. A passed label-relative verifier establishes exclusion with respect to D; the independently annotated measurements determine how far that protection extends to T.

8.3 What does synthetic evolution improve?

The provider-adapted Aug-PE study evaluates distributional coverage when the output consists of generated candidates rather than source sentences. Its endpoint is mean nearest-cosine coverage from each input document to eight final samples versus the first eight initial samples. We schedule 30 fresh generation replicates per corpus on the fixed 40-document cohorts. Successful-run mean differences are positive for every corpus (Table 4), ranging from 0.007 for ai4privacy to 0.120 for single-topic text.

The adaptation uses Copilot SDK 1.0.13, `gpt-5.4-mini`, hashing embeddings, ten histogram rounds, eight retained samples, population multiplier two, and no lookahead. Local Gaussian accounting uses nominal $\varepsilon = 4$, $\delta = 10^{-5}$ and record-add/remove sensitivity one. Model prompts contain only a public genre or previous synthetic candidates; private embeddings and votes remain local. These provider and scale choices define the evaluated adaptation.

The fixed schedule completes 229 cells and records 11 failures across 2329 provider attempts. The generation replicate is the statistical unit. Complete-case 95% intervals use 20,000 replicate bootstrap resamples; the analysis also uses two-sided sign flips and Holm adjustment across eight corpora (Appendix C). Complete-case inference is exploratory. Scheduled-mean bounds allow each unavailable difference to lie in $[-1, 1]$, retaining every failed cell without outcome-dependent replacement.

The scheduled-mean bounds remain positive for five corpora and span zero for ai4privacy, OpenReview, and Yelp. Thus the sign of the finite scheduled mean is robust to these worst-case missing outcomes in five corpora. The contrast measures the combined effect of selection and additional generation; without a matched random-selection evolution control, it does not isolate their contributions. Together with the headroom counterexample, this result motivates evaluating evolution at its final output rather than inferring its potential from the initial candidate family's utility.

9 Discussion and Limitations

Implications for summarization systems. The analysis identifies three different design decisions. The feasible family and objective determine the headroom available to selection. Detection and provenance determine which source spans can be excluded. The release universe and observation process determine whether selection privacy extends to the output. A larger protection radius acts on the second decision; it neither restores missing output support nor changes the detector's recall. Public-output selection instead addresses the support boundary by changing what can be released.

Empirical scope. The core results concern 40 documents per corpus, a 24-sentence prefix, and lexical hashing; a fixed similarity threshold permits different graph densities. Utility is measured by the stated objectives, rather than comprehensive summary quality. Generation replicates share fixed input cohorts, so their uncertainty concerns fresh generations on those cohorts rather than new documents. The medical evidence covers biomedical abstracts and published clinical cases, with broader annotation targets in E3C; patient-disjoint EHR validity and model-training decontamination are unestablished. The synthetic headroom diagnostic measures decision behavior, not computation savings or real-corpus deployment benefit.

Privacy scope. The graph-level adjacency assumes stable survivor quantities and trusted contributor attribution. The private-text extractor has data-dependent support and therefore falls outside the public-universe theorem. The public-label construction satisfies its stated mathematical model; host isolation, publication behavior, and cross-session accounting require operational evidence. Appendix A makes these conditions explicit and gives a quantitative transcript-robustness bound. Fixed benchmark seeds and released diagnostics serve reproducibility, not operationally private release.

10 Conclusion

We establish a framework for evaluating utility, sensitive-content exclusion, and privacy as distinct claims in graph-based summarization. Tight headroom analysis characterizes the gain available to selection; coverage and provenance yield source-exclusion guarantees; and fixed public support with bounded score sensitivity yields conditional selection privacy. Counterexamples prove where these arguments separate. The empirical evidence connects those boundaries to observed selection gains, detector-coverage gaps, and suppression costs. The resulting methodology makes release claims precise: each guarantee is supported by the assumptions and measurements that actually determine it.

AI Use Statement

Generative AI tools assisted with research-method design, mathematical claim and proof development, implementation, experimental configuration, data-processing code, result interpretation, manuscript drafting and editing, and automated critique. GitHub Copilot SDK with `gpt-5.4-mini` generated the synthetic candidates in the reported provider-adapted study. Local GPT-2 was used in separate development probes, outside the reported evaluation. Automated critique was AI-assisted rather than human review. Recorded model outputs and replay checks are retained with the experimental artifacts.

Ethics Statement

Sensitive-content evaluation warrants care even with public datasets. External text remains in local caches and is not redistributed with aggregate results. The disease intervention is synthetic and represents no patient's diagnosis. Medical inference is entirely local, with no clinical text sent to hosted models. Hosted generation receives only public prompts and synthetic candidates. Dataset and model terms are recorded with their provenance. The reported coverage gaps rule out using this research prototype as a clinical de-identification service.

Reproducibility Statement

The experimental artifacts include locked protocols, dataset revisions and checksums, document IDs, source hashes, per-seed records, and aggregate statistics. Tables and figures use recorded outputs; successful hosted outputs support exact computational replay. Proofs state their assumptions explicitly. Appendices B–D document preprocessing and model evaluation, Appendix G details missing uniform draws, and Appendix L gives reproduction commands.

References

F. McSherry and K. Talwar. Mechanism design via differential privacy. In *FOCS*, 2007. doi:10.1109/FOCS.2007.41.

R. McKenna and D. Sheldon. Permute-and-flip: A new mechanism for differentially private selection. arXiv:2010.12603, 2020.

Z. Ding, D. Kifer, S. M. Saghaian N. E., T. Steinke, Y. Wang, Y. Xiao and D. Zhang. The permute-and-flip mechanism is identical to report-noisy-max with exponential noise. arXiv:2105.07260, 2021.

J. Blocki, A. Blum, A. Datta and O. Sheffet. Differentially private data analysis of social networks via restricted sensitivity. arXiv:1208.4586, 2013 revision.

C. Xie, Z. Lin, A. Backurs, S. Gopi, D. Yu, H. A. Inan, H. Nori, H. Jiang, H. Zhang, Y. T. Lee, B. Li and S. Yekhanin. Differentially private synthetic data via foundation model APIs 2: Text. In *ICML*, 2024. arXiv:2403.01749.

R. I. Dogan, R. Leaman and Z. Lu. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, 2014. doi:10.1016/j.jbi.2013.12.006.

d4data. Biomedical NER model card. `https://huggingface.co/d4data/biomedical-ner-all`, checkpoint `015a4050c9ac99722e61c547aa9b4282bcbedc7f`. Apache-2.0.

R. Zanoli, A. Lavelli, D. Verdi do Amarante and D. Toti. Assessment of the E3C corpus for the recognition of disorders in clinical texts. *Natural Language Engineering*, 2023. doi:10.1017/S1351324923000335.

H. Lin and J. Bilmes. A class of submodular functions for document summarization. In *ACL*, pp. 510–520, 2011. `https://aclanthology.org/P11-1052/`.

I. Mironov. On significance of the least significant bits for differential privacy. In *CCS*, 2012.

C. Ilvento. Implementing the exponential mechanism with base-2 differential privacy. arXiv:1912.04222, 2020.

Z. Ding, Y. Wang, G. Wang, D. Zhang and D. Kifer. Detecting violations of differential privacy. In *CCS*, 2018. doi:10.1145/3243734.3243818.

Sumy contributors. Sumy: Automatic text summarizer. `https://github.com/miso-belica/sumy`, revision `7c418d95`. Apache-2.0.