GEPI: DIFFERENTIALLY PRIVATE RELATIONAL EVIDENCE RELEASE FOR LANGUAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Differentially private (DP) retrieval-augmented generation bounds the influence of any one record on the answer, but it inherits the assumption that the generator's evidence is text itself: retrieve first, then control what that text may reveal. This leaves a prior question unasked—what evidence does the answer actually need? We characterize that need as knowledge (a fact or rule), population (a pattern across records), case–population (a population pattern applied to the questioner's conditions), or case (the contents of a designated record), and keep it separate from access permission. Across the benchmark settings of prior private-RAG work, retrieving from a protected corpus does not tell us what evidence the answer needs, or whether it needs private evidence at all. When it does, what must be preserved? We find that *relationships* among facts are key: representations that keep field names and within-record links retain the required evidence for user queries, whereas the same terms and values without those links lose most of it. We therefore introduce GEPI (Global Evidence for Private Statistical Inference), which separates private evidence release from language generation. GEPI encodes the relationships behind a declared set of measurements as bounded document-level statistics and releases one corpus-level synopsis under document-level DP. The language model answers from this release alone, so further supported questions cost no additional privacy budget and never revisit the documents. Across legal and medical population workloads with supplied measurement bindings and two answer generators, GEPI achieves the highest mean reference alignment among the evaluated methods at every finite-budget label. It also improves numerical fidelity, while the specified record-completion attack shows no observed AUC increase over auxiliary information. A relational DP release thus preserves the evidence answers need without exposing record-level context to the generator, and one release serves the whole workload.

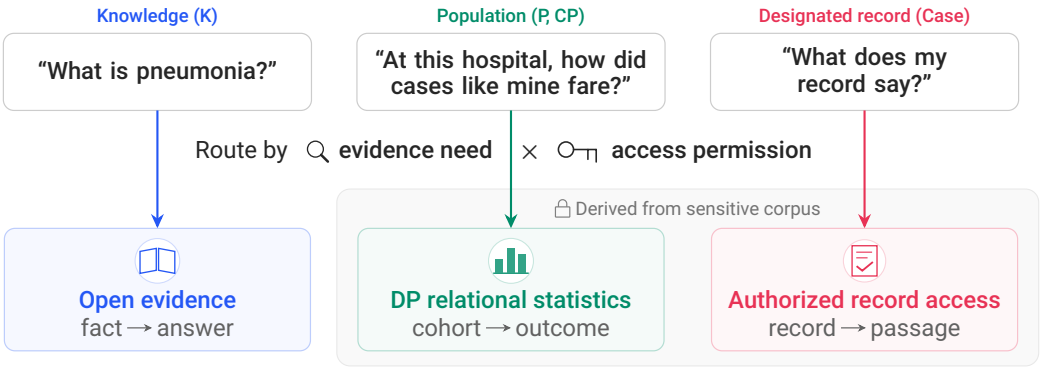


Figure 1: Evidence requirements and access paths. Open sources can support public facts; approved DP releases can support population measurements; designated-record requests require access to that record. These are conceptual routes, conditional on evidence sufficiency and permission.

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.

1 INTRODUCTION

Retrieval-augmented generation (RAG) answers questions by passing retrieved documents to a language model. Medical records and legal files, however, bundle many facts about one individual case, while a question may require only a particular fact or a pattern across cases. Passing entire records to the generator therefore makes private details available beyond what the answer requires. Differentially private (DP) RAG methods limit the influence of individual records by aggregating next-token predictions across retrieved records, releasing DP keyword sets, or replacing the corpus with synthetic text (Grislain, 2025; Koga et al., 2025; Tang et al., 2026; Mori et al., 2026). All of them treat document content, whether original, distilled, or synthetic, as the evidence to privatize, without first asking whether it is the evidence the answer needs.

That premise leaves a prior question unasked: *what evidence does the answer actually need?* A user may ask what a legal principle means, how patients with a condition typically respond to a treatment, how cases like theirs tend to end, or what their own judgment says. Only the last request needs an individual record. We name these four needs **knowledge** (K), **population** (P), **case–population** (CP), and **Case**, and we keep evidence need apart from access permission: the labels say what the answer needs, not who may read a source (Figure 1). When a request needs only population evidence, retrieving individual records grants more access than the task requires; the system can aggregate first and expose only the result. The benchmark settings of prior private-RAG work make this concrete: clinical questions call for population patterns, public pages answer most factual queries, and only the contract questions need the text of one designated record (Section 2). A fidelity bound (Proposition 1) completes the picture: when an answer must change with one protected document, DP itself forces an irreducible error on it, no matter how the mechanism is designed.

Once the required evidence is identified, the next question is how to represent it. A medical record states that a patient had a diagnosis and later received a treatment; counting diagnosis mentions and treatment mentions separately loses that the treatment followed that diagnosis in the same patient. Audits on ClinSQL and Korean LegalQA (Shen et al., 2026; Jeon, 2021) bear this out: the same terms and values, stripped of the links between them, lose most of the required evidence (Section 2). No existing private-RAG interface guarantees that those links, or the counting unit and denominator behind them, survive.

We therefore introduce **GEPI** (Global Evidence for Private Statistical Inference), which releases those relations directly. Before any protected access, the data holder declares which relations will be counted, for instance, among drunk-driving cases with a prior conviction, how many ended in a prison sentence. GEPI maps each document to bounded contributions to those counts, keeps unknown states explicit, and aggregates the contributions into one document-level DP synopsis. The generator answers from the release and the request, told only which released measurements the request uses, and never reopens a protected document, so answering again spends no further privacy budget. Requests for one designated record lie outside this release and require separately authorized access.

We evaluate GEPI against synthetic-release, per-query, and raw-context private-RAG baselines on the same requests, generators, and reference criteria in legal and medical settings. Across both domains, two answer generators, and every evaluated finite privacy budget, GEPI achieves the highest mean final-answer quality among all private baselines. It carries its released statistics into the answer with a fraction of the synthetic baselines' numerical error and supports population answers without increasing a record-completion attacker's advantage over auxiliary information. The raw-context comparator shows the trade-off GEPI avoids: removing clipping reveals record-specific information, while enforcing its transcript budget suppresses that signal and the population answer with it.

Our contributions are threefold. First, we reformulate the private-RAG problem: a taxonomy that separates the evidence a request needs from the access it is granted, audits showing that prior benchmark settings conflate the two, and a fidelity bound on document-specific answers under DP. Second, we show that relations among facts carry most of the evidence clinical and legal questions need, which motivates relational statistics over document-derived text. Third, we introduce GEPI, which releases reusable document-level DP relational statistics and answers from them rather than from retrieved private documents. Across four domain–generator settings, GEPI achieves the highest mean reference alignment at all three evaluated finite-budget labels, with numerical-fidelity and record-completion checks characterizing its released evidence.

2 WHAT EVIDENCE DOES PRIVATE RAG NEED?

Retrieving a protected document does not show that its raw text is needed to answer, and suppressing raw text does not show that the needed evidence survives. Three audits separate these questions, and a fidelity bound shows what differential privacy permits when an answer depends on one document.

2.1 EVIDENCE NEED IS DISTINCT FROM ACCESS PERMISSION

We label each request by the evidence it needs. **Knowledge (K)**: a fact or rule. **Population (P)**: a pattern across records. **Case–population (CP)**: that pattern under the questioner's own conditions. **Case**: what one designated record says. The labels describe evidence, not permission: an internal rule is still K, and a public patient record is still Case. Each need can be met by open evidence, by an approved DP release, or by an authorized source; Table 24 (Appendix G.4) lists the routes.

Finding 1: retrieval settings span different evidence needs. Benchmarks that prior private-RAG methods treat alike ask for different evidence (Figure 2a). In the reconstructed Grislain medical setting, every question is CP: a patient asks how people like them fared (Grislain, 2025). ChatDoctor consultations, used by DPVoteRAG, almost always carry such a population component alongside the individual complaint (Koga et al., 2025). SearchQA clues, used by DP-SynRAG, almost never do; they are knowledge questions, with a single population exception in 408 (Mori et al., 2026). ContractNLI sits at the other extreme: each question asks what one designated NDA says, so every request is Case (Koreeda & Manning, 2021; Pipitone & Houir Alami, 2024). Compound requests are labeled by component and counted once. Two corpora served by the same token-aggregation interface can therefore need population evidence nearly every time or essentially never. Sampling and labeling procedures are in Appendix G.

2.2 OPEN SUPPORT AND RELATIONAL EVIDENCE

Finding 2: public evidence fully supports answers to two thirds of audited requests. Across two audit stages covering all 408 SearchQA requests, admitted public pages yield benchmark-aligned answers for about six in seven (351/408; Figure 2c), without touching the protected corpus. A support check that never sees the benchmark answer rates two thirds of all requests fully supported by their citations (272/408), 77 partially, and only two unsupported. When open evidence suffices, the interface can answer without protected access or privacy budget. Appendix G details the screens and the two-stage protocol.

Finding 3: preserving terms does not preserve their relationships. Population evidence lives in relations, not in the words. For each of 340 ClinSQL and 400 Korean LegalQA requests (Shen et al., 2026; Jeon, 2021), we list the evidence items its reference SQL or prose requires, build a typed representation, and check whether every item is still recoverable. A *linked* representation, which keeps field names and the links within a record, retains all required evidence for roughly nine in ten assessable requests in both corpora. A control that keeps exactly the same terms, values, and authority strings but drops the field names and links loses most of them: in the clinical audit fewer than one request in seven survives intact, and in the legal audit fewer than one in three (Figure 2b). What disappears is always the same kind of thing: which patient the diagnosis belongs to, whether treatment followed it, which cohort a case falls in. Cohort membership, attribution, and condition–outcome order carry evidence that isolated mentions cannot (Appendix G).

2.3 A FIDELITY LIMIT FOR DOCUMENT-SPECIFIC ANSWERS

A Case request asks what one designated record says. When that record is protected and is the sole source of a requested finding, a correct answer reports the finding when the record is present and indicates insufficient evidence when it is absent. For such Case requests, fidelity therefore requires the answer to change with a single protected document. Document-level DP limits how much the answer distribution may change between these two corpora (Section 3), creating a tension between document-level privacy and Case-answer fidelity.

We quantify this tension by comparing the mechanism's answer distribution with a reference distribution supported on correct answers for each corpus. The public request and auxiliary information are held fixed, and answers are interpreted under the same semantic rule on both corpora. A reference distribution assigns all its mass to answers accepted by the task's correctness rule, reducing to a

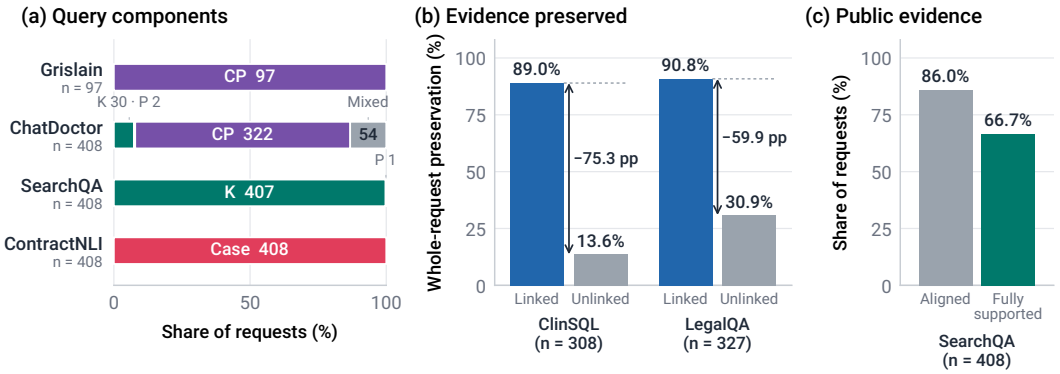


Figure 2: (a) Evidence needs across settings used in prior private-RAG evaluations. Population components dominate the medical settings (Grislain, ChatDoctor), are essentially absent from SearchQA, and ContractNLI is Case by construction. "Mixed" counts a compound request once. (b) Whole-request evidence preservation on ClinSQL and LegalQA. The linked representation keeps nearly every request intact; removing field names and within-record links, while keeping the same terms and values, loses most of them. Arrows give paired differences. (c) Public-evidence answers across two audit stages on 408 SearchQA requests. Fully supported answers are a subset of benchmark-aligned answers; both bars use all 408 requests as the denominator. The stages used different search procedures (Appendix G).

point mass when the semantic answer is unique. Total variation, $\mathrm{TV}(P, Q) = \sup_E |P(E) - Q(E)|$, measures the discrepancy in probabilities of semantic answer events rather than differences in wording.

Proposition 1 (Case-answer fidelity). *Let $D_1 = D_0 \cup \{d^\star\}$ with $d^\star \notin D_0$. For each $i \in \{0, 1\}$, let P_i be a reference distribution supported on answers accepted by the task's correctness rule on corpus D_i, and let Q_i be the complete-answer distribution of the same (ϵ, δ)-DP mechanism on that corpus. If P_0 and P_1 have disjoint support, $\epsilon \geq 0$ and $0 \leq \delta < 1$, then*

$$\max_{i \in \{0,1\}} \mathrm{TV}(P_i, Q_i) \geq \frac{1 - \delta}{1 + e^\epsilon}. \tag{1}$$

In the example above, reporting the finding and reporting insufficient evidence are distinct correctness requirements, giving disjoint reference supports. At $\epsilon = 1$ and negligible δ, the mechanism must therefore incur a TV discrepancy of at least approximately 0.27 on one of the two corpora. Its scope depends on the required answer's dependence on the protected document, rather than on the Case label alone: if public information or other records also support the finding, the disjoint-support assumption may no longer hold. Appendix D.1 gives the proof and the overlapping-reference and Gaussian-release versions.

2.4 Implications for private RAG and GEPI

The audits and the fidelity bound together motivate: route each request by the evidence it needs. Use open evidence when it suffices; use separately authorized source access when the answer requires a designated protected record; and, when releasing protected population evidence, preserve cohort, denominator, outcome, and within-record relations. Existing private-RAG interfaces fall short on the last point. Token or logit aggregation can answer K, P, and CP requests, but agreement among document-conditioned predictions does not specify a corpus denominator (Grislain, 2025; Koga et al., 2025); keyword marginals convey cues without the relations (Tang et al., 2026); synthetic corpora are reusable, but their population fidelity depends on keeping the counting unit and linked fields (Mori et al., 2026).

GEPI makes those measurements the object of release. It encodes declared cohort–outcome relations as bounded document-level contributions, retains missing evidence as explicit unknown states, and releases their aggregate under document-level DP. The generator then answers supported questions from this single synopsis without revisiting the protected documents.

3 PROBLEM SETUP

Protected unit and public transcript. Write $D \simeq D'$ when the datasets differ by adding or removing one document: a case bundle in our LegalM criminal-judgment corpus or a medical consultation case. This is document-level, not person-level, protection. A mechanism M is (ϵ, δ)-differentially private if, for all $D \simeq D'$ and every measurable S,

$$\Pr[M(D) \in S] \leq e^{\epsilon} \Pr[M(D') \in S] + \delta. \tag{2}$$

Here δ is additive slack (Dwork & Roth, 2014). GEPI uses the central model: document extraction and aggregation remain inside the trusted data holder's boundary. The generator receives the release, the request, and its public measurement bindings (which declared measurements the request uses; defined below), not protected documents. The public transcript includes the fixed public workload and bindings, the release, and answers computed from it. Section 4 bounds the joint contribution across measurements and calibrates the release noise; empirical attacks test a specified inference task separately.

Declared measurements. GEPI fixes what it will measure before reading any protected document. These measurements form a public workload $\mathcal{W}$: a finite set of *measurement contracts*. Each contract specifies which documents belong to a cohort, which outcome to determine within each document, the measurement parameters, and how to handle missing evidence. For example, a contract may measure whether treatment began within two months of a specified diagnosis in the same record. If the timing information is missing, the outcome is recorded as unknown rather than negative.

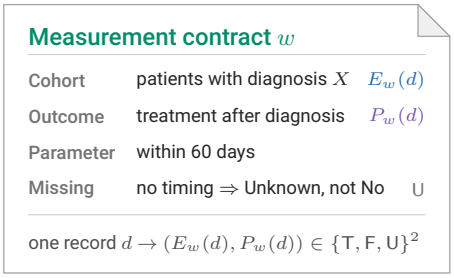


Figure 3: A measurement contract. It names a cohort, a within-document outcome, its parameters, and a missingness rule; every document is reduced to a bounded state under it (Section 4). Contracts are fixed before any protected document is read.

A natural-language request may require one or more of these measurements, and different requests may use the same measurement. The request's public *measurement bindings* identify the contracts whose released results provide its evidence. Thus, $\mathcal{W}$ defines the available measurements, while the bindings connect each request to the measurements it needs. In our evaluation, task construction supplies these bindings so that all methods are compared on the same target measurements. We evaluate evidence release and answer generation under these supplied bindings; automatic matching of free-form requests to contracts is outside the evaluation.

4 GEPI: RELEASING RELATIONAL EVIDENCE

Section 2 identifies three requirements for population evidence: preserving within-record relations, counting them over a defined cohort, and distinguishing missing evidence from negative outcomes. Passing retrieved text to a generator does not by itself ensure that these requirements are met. GEPI addresses them by making the required population measurements explicit and releasing their results before language generation.

Before reading any protected document, the data holder fixes the cohort definitions, outcome relations, and missingness rules. Each document is then mapped to bounded contributions with explicit unknown states inside the trusted boundary. GEPI aggregates these contributions and adds Gaussian noise once to produce a reusable synopsis. The generator answers supported questions from this synopsis without reopening the documents. Because these answers are post-processing of the same release, reuse incurs no additional privacy loss. Algorithm 1 summarizes the procedure; implementation details, calibrations, and prompts appear in Appendices I and I.5.

4.1 DOCUMENT STATES AND BOUNDED ENCODING

Population counts are useful only if cohort conditions and outcomes refer to the same records. GEPI evaluates attribution and temporal conditions within each document before aggregation. A fixed local extractor supplies evidence; deterministic rules assign true, false or unknown states. Missing

Algorithm 1 GEPI: one evidence release, answers to original questions

Require: Documents D; public measurements, recipes, and request bindings; fixed local extractor; target (ϵ, δ); privacy ledger; fixed readout policy

Ensure: A reusable noisy synopsis $\widetilde{H}$ and answers derived from it

A. Compile public specifications before the private mechanism

Freeze cohort/outcome definitions, unknown rules, coordinate layout, and public recipes

Bound each document's contribution, $\|x(d)\|_2 \le \sqrt{B}$, and calibrate σ to (ϵ, δ)

★ B. Measure relations inside each document

$H \leftarrow 0$

for each document $d \in D$ **do**

    Extract local evidence; evaluate conditions, retaining unknown states

    Encode bounded contribution $x(d)$; $H \leftarrow H + x(d)$

end for

C. Account and publish once

Reserve the Gaussian privacy charge in the ledger; stop if unavailable

$\widetilde{H} \leftarrow H + Z$, $Z \sim \mathcal{N}(0, \sigma^2 I)$ with fresh randomness

Commit the charge; publish and persist $\widetilde{H}$ and public recipes for reuse

D. Answer without reopening documents

for each original question q and user context c **do**

    Resolve the supplied public bindings against the released recipes; absent recipe ⇒ `UNSUPPORTED`

    Read the bound counts from $\widetilde{H}$, zero-clip them, and form the interval of Eq. (6) (or apply the precision policy)

    Generate from unchanged q, c, the released evidence, and its limitations

    Validate evidence use and output schema; otherwise return a public fallback

end for

Bindings are supplied by the workload adapter. Reuse adds no privacy charge.

evidence, unresolved conditions and invalid schemas remain unknown rather than becoming negative outcomes. Unknown eligibility and unknown outcomes are counted separately, keeping the denominator explicit (Figure 3).

GEPI groups measurements with shared structure into *families*. For example, several outcomes for the same cohort can share a denominator while retaining separate positive and unknown counts. Sharing avoids separately releasing statistics for each answer, while the fixed layout specifies which relations are available; any requested conjunction must be represented before release. Each document belongs to one predeclared partition and produces a binary vector $x(d)$ with at most B active coordinates. Thus $H(D) = \sum_d x(d)$ aggregates the corpus and $\|x(d)\|_2 \le \sqrt{B}$ bounds one document's influence. Sensitivity depends on this joint contribution, not the number of later answers. Public recipes reconstruct the declared counts from these coordinates; the released evidence contains aggregate measurements, not passages or document-level relations. Appendix D.4 gives the state definitions, count formulas and family layouts.

4.2 One private release

The coordinate universe, extractor, and recipes are fixed from public information before protected access, as the guarantee requires. Adding or removing one document changes H by $x(d)$, so

$$\Delta_2(H) := \sup_{D \simeq D'} \|H(D) - H(D')\|_2 \le \sqrt{B}, \tag{3}$$

$$\widetilde{H} = H(D) + Z, \quad Z \sim \mathcal{N}(0, \sigma^2 I). \tag{4}$$

Noise is added to every coordinate, including zero counts. With $\mu = \sqrt{B}/\sigma$, the analytic Gaussian calibration selects σ such that

$$\Phi\left(\frac{\mu}{2} - \frac{\epsilon}{\mu}\right) - e^{\epsilon}\Phi\left(-\frac{\mu}{2} - \frac{\epsilon}{\mu}\right) \le \delta. \tag{5}$$

This yields document-level (ϵ, δ)-DP (Balle & Wang, 2018); Proposition 2 states the joint release-and-answer guarantee. Every answer computed from the saved release and public information is post-processing. Supported reuse therefore incurs no additional privacy loss; a new noisy draw or a new private measurement requires accounting. Each release's noise profile and guarantee appear in Appendix D.7.

4.3 PUBLIC READOUT AND GENERATION

The generator never sees a document. It receives the original request, user context, and supplied measurement bindings; a binding names a public recipe, and an absent recipe returns `UNSUPPORTED`, not zero. The evaluation holds these bindings fixed, so it measures release fidelity and generation rather than open-ended target selection.

For interval tasks, the main experiments read the two endpoints of the reported interval, its lower and upper limits, directly from the release. Let a, n, u, u_E be zero-clipped estimates of the counts of eligible documents with the outcome, without it, with unknown outcome, and documents with unknown eligibility, respectively. The reported interval is

$$\left[\frac{a}{m}, \frac{a+u+u_E}{m}\right], \qquad m = a+n+u+u_E, \qquad m=0 \mapsto [0,1]. \tag{6}$$

On true counts, this interval contains the known-cohort interval under every eligibility assignment of the u_E documents. On noisy counts it is a descriptive readout reported with the noise scale, not a confidence interval. The language model explains this evidence for the cohort named by the binding without using general knowledge to invent a missing corpus frequency.

An optional noise-aware precision policy can instead withhold unstable ratio answers (Appendix D.8); it is not the direct readout used in the main utility table. Appendix I.3 specifies the release and answer protocol.

5 EXPERIMENTS

We test answer utility, preservation of numerical evidence, and record-completion advantage from released outputs.

5.1 EXPERIMENTAL SETUP

Tasks. We evaluate evidence release and answer generation for declared population measurements on LegalM (our Korean criminal-judgment corpus) and LeakRAG medical consultations (Joo et al., 2026a;b). Every request needs population evidence about the corpus: a legal request appends a population question to an ordinary consultation, and a medical request asks for a cohort proportion as an identified interval. The measurement, which each request needs, is supplied with it, so all methods answer the same statistical demand. Figure 4 summarizes the setting; construction and protocol history are in Appendices I.5 and A.

Experimental setting

	Legal	Medical
Corpus	LegalM, 3,000	LeakRAG, 1,200
Request	consultation + population question	cohort proportion, as an interval
Requests	300 (245 population, 55 consultation-only)	300
Reference	exact aggregate of frozen document states, evaluator-only	
Generators	GPT-5.4-mini, Qwen3.8-27B	
Methods	GEPI, Tab-PE, Aug-PE, per-query count DP, raw-context token DP; public-only at $\epsilon = 0$	
Budgets	$\epsilon \in \{1, 4, 8, \infty\}$	
Score	reference alignment, 1–5	

Figure 4: Experimental setting. Within a block (domain $\times$ generator), every method answers the same 300 requests under the same scoring contract. The 55 consultation-only controls receive identical outputs from every method.

Methods and scoring. GEPI is compared with representative baselines: Tab-PE synthetic tables (Tran et al., 2026), Aug-PE synthetic text (Xie et al., 2024), Grislain-style raw-context token DP (Grislain, 2025), and per-query count DP; public-only generation receives no release. Tab-PE and count DP share GEPI's frozen eligibility and outcome states, holding document interpretation fixed while changing the release and readout. Population references are common exact aggregates of these states, independent of each method's noisy output. The primary score is complete-answer reference alignment (1–5), assigned against the exact aggregate by a GPT-5.4-mini judge with a domain-specific rubric. Appendix C.4 reports a descriptive comparison with a practicing lawyer's blind scores on legal answers. Within a block, every method shares the requests, generator and judge; means use scored requests. Rubrics and scored examples are in Appendices C and B.

Table 1: Reference alignment (1–5, ↑): scored mean ± conservative 95% cluster-bootstrap half-width (legal: population statistic; medical: geometry family). Each block uses 300 common requests; Table 18 gives scored counts and completion bounds. Bold: highest column mean. Budget labels are nominal (guarantees: Appendix D.7); $\epsilon = 0$ is shared public-only generation, and $\epsilon = \infty$ denotes method-specific non-private conditions.

Setting	Method	$\epsilon = 0$	$\epsilon = 1$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = \infty$
⚖ Legal GPT-5.4-mini	**GEPI** (ours)	2.997 ± 0.08	$\mathbf{3.121} \pm 0.10$	$\mathbf{3.322} \pm 0.10$	$\mathbf{3.386} \pm 0.09$	3.974 ± 0.09
	Tab-PE		2.605 ± 0.11	2.621 ± 0.11	2.682 ± 0.11	3.970 ± 0.09
	Aug-PE		3.030 ± 0.09	2.967 ± 0.09	2.987 ± 0.09	2.983 ± 0.09
	Per-query count DP		2.789 ± 0.12	3.147 ± 0.12	3.385 ± 0.12	$\mathbf{4.522} \pm 0.07$
	Raw-context token DP		3.104 ± 0.08	3.030 ± 0.09	3.084 ± 0.09	3.084 ± 0.09
⚖ Legal Qwen3.8-27B	**GEPI** (ours)	2.664 ± 0.08	$\mathbf{3.060} \pm 0.10$	$\mathbf{2.949} \pm 0.12$	$\mathbf{3.127} \pm 0.11$	$\mathbf{3.909} \pm 0.08$
	Tab-PE		2.155 ± 0.09	2.215 ± 0.09	2.239 ± 0.09	3.841 ± 0.08
	Aug-PE		2.655 ± 0.08	2.615 ± 0.08	2.646 ± 0.08	2.741 ± 0.08
	Per-query count DP		2.561 ± 0.10	2.738 ± 0.10	2.976 ± 0.11	3.900 ± 0.08
	Raw-context token DP		2.748 ± 0.08	2.722 ± 0.08	2.702 ± 0.08	2.679 ± 0.08
Medical GPT-5.4-mini	**GEPI** (ours)	2.000 ± 0.00	$\mathbf{2.663} \pm 0.09$	$\mathbf{3.360} \pm 0.10$	$\mathbf{3.662} \pm 0.11$	4.343 ± 0.11
	Tab-PE		2.250 ± 0.07	2.430 ± 0.08	2.460 ± 0.09	2.361 ± 0.09
	Aug-PE		2.083 ± 0.05	2.160 ± 0.07	2.153 ± 0.07	2.173 ± 0.07
	Per-query count DP		1.933 ± 0.08	2.157 ± 0.09	2.473 ± 0.12	$\mathbf{4.870} \pm 0.06$
	Raw-context token DP		1.973 ± 0.03	1.956 ± 0.04	1.963 ± 0.03	1.899 ± 0.05
Medical Qwen3.8-27B	**GEPI** (ours)	1.997 ± 0.01	$\mathbf{2.562} \pm 0.08$	$\mathbf{3.160} \pm 0.08$	$\mathbf{3.211} \pm 0.08$	3.563 ± 0.10
	Tab-PE		2.137 ± 0.07	2.277 ± 0.07	2.304 ± 0.08	2.328 ± 0.08
	Aug-PE		1.961 ± 0.07	2.007 ± 0.07	2.011 ± 0.07	2.115 ± 0.07
	Per-query count DP		1.941 ± 0.07	2.125 ± 0.08	2.332 ± 0.11	$\mathbf{3.611} \pm 0.10$
	Raw-context token DP		1.970 ± 0.03	1.940 ± 0.04	1.898 ± 0.04	1.907 ± 0.04

5.2 Answer Utility on Corpus-Grounded Benchmarks

Does relational release improve the requested answers? GEPI has the highest mean reference alignment among all private baselines at every finite-budget label ($\epsilon \in \{1, 4, 8\}$) in all four domain–generator settings (Table 1). In medical, even its $\epsilon = 1$ means exceed the synthetic baselines' non-private means. Appendix F.6 gives paired comparisons and unresolved-output sensitivity. The revised Legal–Qwen $\epsilon = 1$ reader also improves on its predecessor on 60 pairs excluded from reader development and common-300 ($+0.533 \pm 0.213$, cohort-cluster 95% CI; Appendix I.5).

Why share one synopsis? Count DP uses the same frozen states but releases query-specific counts. GEPI has higher finite-label means; exact counts lead in three blocks without noise. In medical GPT at $(\epsilon, \delta) = (4, 10^{-5})$, using a common evidence format and partition-aware count accounting gives GEPI 3.360 versus count DP 2.497 (paired $+0.863 \pm 0.120$, cluster 95% CI). In 200 noise simulations at the same medical setting, GEPI has lower roster-mean maximum-endpoint error than partition-aware count DP in every repetition (0.079 versus 0.223 on average), using a common interval decoder (Appendix E). Raw-context results characterize its tested adapter: its medical population-evidence gate fails even without clipping (Yu et al., 2026).

5.3 Numerical Evidence Fidelity

To check numerical content independently of the primary judge, a score- and method-blind audit compares interval spans in earlier frozen medical outputs with exact references at $\epsilon = 4$ (30 common targets per method and generator; 240 outputs). Conditional on a retained interval, GEPI's mean maximum-endpoint error is 0.078 (GPT) and 0.088 (Qwen), below 40% of either synthetic baseline's error, with higher or equal parser yield (Table 14). Appendix F.1 gives the extraction protocol; Appendices F.2 and F.3 add full-roster tolerance coverage and a controlled test of information lost when relations are removed.

5.4 Record-Completion Attack on Released Evidence

We test whether released evidence helps an attacker infer whether a target legal consultation states prior criminal punishment. The attacker receives de-identified incident phrases, an event class, and

Table 2: Legal record-completion attack ($N = 87$). Δ AUC is relative to auxiliary-only on common resolved targets; $\pm$ denotes conservative 95% bootstrap half-widths (Appendix H). BA uses all 87 targets and counts the 4/1/1 unresolved raw-context outputs (transcript/per-query/∞) as correct attacker decisions; counting them as errors gives 48.1/50.0/50.0. Lower values indicate less attack success ($\downarrow$). GEPI uses one $(4, 1/3000)$ release; this is not a matched-DP comparison. †Historical public-seed artifacts. §Pure $\epsilon = 4$/query composes to $\epsilon \leq 348$ over 87 queries, unlike the transcript-level $\epsilon \approx 4$ condition.

Release	Method	Δ within-cohort AUC (pp) $\downarrow$	BA (%) $\downarrow$
None	Auxiliary only	0.0 (reference)	50.0
DP release	**GEPI** (ours)	$+0.0 \pm 3.7$	54.1
	Tab-PE†	$+1.1 \pm 2.6$	52.9
	Aug-PE†	$+4.7 \pm 4.7$	51.5
Raw-context token DP	Transcript $\epsilon \approx 4$	$+0.0 \pm 0.0$	52.9
	Per-query $\epsilon = 4^{\S}$	$+1.1 \pm 2.1$	51.5
	Matched $\epsilon = \infty$	$+6.3 \pm 4.8$	51.5

a broad region; the prediction concerns what the consultation records. Every condition gives the same probability reader these clues, with or without one method's public output, and we report the change in within-cohort AUC over the clues alone on 87 targets. A non-private explicit-evidence control yields 80/82 correct predictions, confirming reader sensitivity on the available controls (Appendix H).

GEPI's estimated within-cohort AUC change is $+0.0 \pm 3.7$ percentage points (Table 2). For the raw-context adapter, the unclipped endpoint increases AUC by $+6.3$ points, compared with 0.0 under the transcript budget and $+1.1$ under the per-query budget. The latter interval includes zero. Appendix H specifies the reader, the distinct raw-context utility and attack adapters, and the development history.

6 Related Work

Private query and language-model interfaces. GEPI builds on bounded-sensitivity query release, Gaussian calibration and post-processing (Dwork & Roth, 2014; Balle & Wang, 2018). DP-ICL privatizes grouped exemplar predictions or keywords (Wu et al., 2024); private RAG aggregates document-conditioned outputs (Grislain, 2025; Koga et al., 2025), and DP-KSA releases lexical cues (Tang et al., 2026). GEPI's contribution is an evidence interface that releases bounded measurements of cohort denominators and condition–outcome relations.

Synthetic and domain evidence. Aug-PE and DP-SynRAG release reusable synthetic text (Xie et al., 2024; Mori et al., 2026); Tab-PE generates categorical tables (Tran et al., 2026). GEPI also supports reuse, but releases declared statistical coordinates and reconstruction recipes. Clinical text-to-SQL specifies cohort predicates, temporal joins and aggregation (Lee et al., 2022; Shen et al., 2026). LegalBench, ContractNLI and LegalBench-RAG study legal reasoning and retrieval (Guha et al., 2023; Koreeda & Manning, 2021; Pipitone & Houir Alami, 2024). Our audits use ClinSQL, Korean LegalQA (Jeon, 2021) and ContractNLI; population tasks extend LeakRAG medical (Joo et al., 2026a;b) and the separately constructed LegalM corpus.

7 Conclusion

Private RAG should begin with the evidence a question needs. We separate evidence requirements from access permissions and establish a DP fidelity limit for document-specific answers. For population questions, GEPI makes relations among cohort conditions, outcomes and unknown states explicit, then releases their statistics as a single corpus synopsis under document-level DP. Legal and medical evaluations show improved answer alignment and numerical fidelity with this interface. The design separates protected-data processing from language generation: once the synopsis is released, supported requests can be answered without reopening source documents or incurring additional privacy cost. GEPI makes declared population evidence reusable across a workload, placing the choice of what to release at the center of private-RAG design.

AI USE STATEMENT

AI assisted coding, proofreading, figure icons, and consistency checks. Experimental proposal, evaluation, and attack-label adjudication used language models as documented in the protocols; model-assisted labels are distinguished from human judgments. The authors take responsibility for all content and result verification.

REFERENCES

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 394–403. PMLR, 2018. URL https://proceedings.mlr.press/v80/balle18a.html.

Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022. doi: 10.1111/rssb.12454. URL https://arxiv.org/abs/1905.02383.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. doi: 10.1561/0400000042. URL https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf.

Nicolas Grislain. Rag with differential privacy. *arXiv preprint arXiv:2412.19291*, 2025. URL https://arxiv.org/abs/2412.19291v2. Version 2, revised 22 January 2025.

Neel Guha, Julian Nyarko, Daniel E. Ho, et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1915. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/89e44582fd28ddfea1ea4dcb0ebbf4b0-Abstract-Datasets_and_Benchmarks.html.

Heewon Jeon. LegalQA using SentenceKoBART. GitHub repository, 2021. URL https://github.com/haven-jeon/LegalQA.

Seongho Joo, Hyukhun Koh, and Kyomin Jung. PEARL: Differentially private and entropy-aware regulated language generation. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=dSRLxMgwxd. Introduces the LeakRAG benchmark.

Seongho Joo, Hyukhun Koh, and Kyomin Jung. LeakRAG: PEARL financial and medical dataset. Hugging Face Datasets, 2026b. URL https://huggingface.co/datasets/Lyway/LeakRAG/tree/e6b305b5e514568fe284382d6fc46728bf157e30. Accessed 2026-09-17; snapshot `e6b305b5e514568fe284382d6fc46728bf157e30`, last modified 2026-07-05.

Tatsuki Koga, Ruihan Wu, Zhiyuan Zhang, and Kamalika Chaudhuri. Privacy-preserving retrieval-augmented generation with differential privacy, 2025. URL https://arxiv.org/abs/2412.04697v3. Version 3, revised 11 November 2025.

Yuta Koreeda and Christopher Manning. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1907–1919. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.164. URL https://aclanthology.org/2021.findings-emnlp.164/.

Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. EHRSQL: A practical text-to-SQL benchmark for electronic health records. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2301.07695.

Ilya Mironov. Rényi differential privacy. In *IEEE 30th Computer Security Foundations Symposium*, pp. 263–275, 2017. doi: 10.1109/CSF.2017.11. URL https://arxiv.org/abs/1702.07476.

Junki Mori, Kazuya Kakizaki, Taiki Miyagawa, and Jun Sakuma. Differentially private synthetic text generation for retrieval-augmented generation (RAG). In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1216–1235. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.62. URL https://aclanthology.org/2026.findings-acl.62/.

Nicholas Pipitone and Ghita Houir Alami. LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*, 2024. URL https://arxiv.org/abs/2408.10343.

Yifei Shen, Yilun Zhao, Justice Ou, Tinglin Huang, and Arman Cohan. Patient-similarity cohort reasoning in clinical text-to-SQL. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1367–1412. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.64. URL https://aclanthology.org/2026.eacl-long.64/.

Tingting Tang, James Flemings, Yongqin Wang, and Murali Annavaram. Differentially private retrieval-augmented generation, 2026. URL https://arxiv.org/abs/2602.14374.

Toan Tran, Arturs Backurs, Zinan Lin, Victor Reis, Li Xiong, and Sergey Yekhanin. Differentially private synthetic data via APIs 4: Tabular data. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.08259.

Tong Wu, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. Privacy-preserving in-context learning for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=x4OPJ7lHVU.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, Bo Li, and Sergey Yekhanin. Differentially private synthetic data via foundation model APIs 2: Text. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 54531–54560. PMLR, 2024. URL https://proceedings.mlr.press/v235/xie24g.html.

Hyojeong Yu, Hyukhun Koh, Minsung Kim, Yunah Jang, and Kyomin Jung. PRAGMA: Evaluating personalized guidance with memory alignment in lifelong conversations. arXiv preprint arXiv:2609.09664, 2026. URL https://arxiv.org/abs/2609.09664.

GUIDE TO THE APPENDICES

Confidence intervals use $\hat{\theta} \pm h$, with h the larger distance to either bootstrap endpoint, rounded outward: a conservative symmetric envelope, not a standard error. Identification intervals and missing-outcome bounds retain endpoint notation.

Appendix	Question addressed
A Task Construction and Reference Targets	How were the tasks, protected units, and reference answers constructed?
B Scored Answer Examples and Error Analysis	What did each method actually answer, and where did the scores diverge?
C Evaluation Protocols and Judge Assessment	What does the score measure, and how reliable is the judge?
D Fidelity Limits, Formal Properties and Privacy Accounting	Which release and transcript carry the DP guarantee, and what fidelity limit follows from it?
E Baseline Specifications and Controlled Comparisons	What does each baseline comparison isolate?
F Additional Results and Robustness	Do coverage, failures, or representation choices explain the results?
G Evidence-Requirement and Preservation Audits	How were evidence requirements and evidence preservation annotated?
H Record-Completion Evaluation	What did the attacker observe, and what do the controls establish?
I Reproducibility and Protocol Deviations	Where are the execution contracts, deviations, and availability recorded?

A TASK CONSTRUCTION AND REFERENCE TARGETS

GEPI is evaluated on requests that explicitly require evidence about a protected corpus. The construction separates the source records, the measurements defined over them, and the questions used to test a released interface.

A.1 CORPORA AND PROTECTED UNITS

Table 3: Corpora used by the population tasks. Document variants and evaluation requests are not additional independent protected contributions.

	Medical	Legal
Source	LeakRAG medical split	LegalM Korean criminal judgments
Protected units	1,200 original consultation cases	3,000 case bundles
Document rendering	Four variants per case (4,800 files)	Source judgment plus accepted generated consultation Q&A
Population target	Cohort intervals for documented markers	Outcome frequencies in declared offense cohorts
Reference	Exact aggregates of frozen case states	Exact aggregates plus separate consultation prose references

Medical uses the LeakRAG split distributed with PEARL (Joo et al., 2026a;b); it does not treat four variants of a case as four observations. LegalM is a separate domain constructed in the LeakRAG style, with synthetic identifiers and generated consultation questions and answers. Its five offense partitions contain 600 defamation/insult, 400 fraud, 400 sexual-offense, 600 traffic, and 1,000 violent-offense bundles. Of 3,000 bundles, 2,967 have accepted consultation Q&A and 33 are source-only; the latter remain in the protected denominator. The acceptance checks are model-based. Every method uses the same frozen bundle identities.

The medical documents retain the received input snapshot, including an encoding artifact in a subset of files; no silent text repair was applied to the frozen inputs. The detailed byte inventory and earlier 443/500/2,000-bundle construction stages are retained with the input provenance rather than used as current corpus statistics. Redistribution scope is specified in Appendix I.

A.2 From a corpus to a supplied measurement

The private measurement workload defines cohorts, documentary outcomes, and unknown states. Its coordinate layout and document-local mapping are fixed before protected measurement. The request-to-measurement binding is then fixed before answer generation; these are distinct commitments. Medical questions name a complaint partition, documentary marker conditions, and an outcome marker. They request the identified interval of Equation 6, including unknown eligibility and outcomes. They concern documentation in the corpus rather than patient prevalence.

Each legal population request retains a consultation and appends a researcher-authored corpus-frequency question. Public hash ordering within offense strata pairs the statistic with the consultation parent. This holds the statistical demand fixed for comparing evidence interfaces; it does not infer which statistic is most relevant to the consultation text. The parent, target statistic, and cohort are therefore different units. New parents under this workload do not constitute new measurement-family generalization.

Why add a population requirement? The added question makes corpus-specific evidence an explicit, scorable requirement while retaining the legal consultation component. All methods receive the same request, measurement binding and exact statistical reference. This evaluates evidence-to-answer fidelity under a supplied task contract; the separate consultation-only control is in Appendix I.7.

Evaluation rosters. Table 1 uses the same 300 requests for every method and $\epsilon = \infty$ condition within each domain, and the two generators share the roster. Legal includes 245 population requests and 55 consultation-only controls; medical includes 300 population requests. Original development and Qwen pilot parents are excluded. Legal retains 224 parents and replaces 76 after the exposure audit; medical retains 240 and replaces 60. Selection follows frozen partition quotas and hash ordering without selecting on scores. These requests use the declared measurement families; the retained medical stratum's common interface clarification and scoring revisions are documented in Appendix I. Exact ($\epsilon = \infty$) conditions also use 300 requests. Failures remain in the denominator (Table 18).

A.3 Fixed task illustrations

The task examples below illustrate the measurement contract, not a method output or a worked numerical result. We selected one medical and one legal population item before inspecting a release, generated answer, score, or evaluator value. For each domain, we compute SHA-256 of `section5-example-panel-v1|domain|stable-item-id` and use the item with the lexicographically smallest digest: the medical candidate set is the earlier frozen 270-parent roster excluding the original medical development set, and the legal candidate set is the earlier 245 population mappings (consultation-only controls excluded). The selection record, input commitments, and resulting deidentified field tuples are recorded in the reproducibility artifact. These illustration pools precede the common-300 replacement. Their Qwen pilot exposure concerns the earlier pools; those pilot parents are excluded from the current Table 1 rosters.

The medical illustration is the selected task with partition `chest_or_epigastric_discomfort`, conditions `labs=yes` and `medication=no`, and outcome `referral`. The legal illustration uses the selected `sexual_offense` cohort and `acquittal_or_dismissal` outcome. Neither these examples nor their selection artifact contains a source document, consultation text, stable case identifier, exact aggregate, model answer, or judge score.

A.4 Reference construction and expert review

Population references are evaluator-only exact aggregates of the same frozen states being released. This supplies a method-independent statistical target: matching a method's own noisy output is not the scoring criterion. Legal consultation components additionally use generated prose references. The numerical target establishes fidelity to the representation; it is not an independently verified clinical or legal answer.

A practising Korean criminal lawyer reviewed 49 selected generated consultation documents from the initial 443-document subset: 34 had identified defects, 10 passed, and five had no verdict. The issues included procedural bars, doctrine, and offense elements. This selected review characterizes

the prose references and does not estimate corpus-wide legal correctness. It is distinct from the 37-answer scoring assessment in Appendix C.4. Development detectors and a revised corpus not used in the reported evaluation remain in the historical artifact.

Source-gold controls. Source-gold controls on public medical records and ContractNLI NDAs test document-local extraction on those datasets (Appendix I.4). Main-corpus references remain the frozen-state aggregates and generated prose described above.

B Scored Answer Examples and Error Analysis

These examples use saved GPT answers and original scores from earlier campaigns, not the replacement Table 1. Selection was post hoc: GEPI scored 5 and the three controls at most 3, with a fixed hash tie-break among qualifying targets. A contrasting case is also shown. This illustrative selection does not estimate average performance; no answer was regenerated or rescored.

Only the public population subquestion and relevant answer excerpts are shown; private consultation text and evaluator-only exact counts are omitted. Legal excerpts are author translations from Korean; medical answer excerpts are verbatim. Original whole-answer scores are retained even when only the population component is displayed. The provenance index records the source packets, answers, judgments, release identities and selection rule.

B.1 Legal: a corpus frequency alongside a consultation

Request and reference. The added question asks how common a fine is among traffic/drunk-driving cases with victim settlement, how UNKNOWN outcomes limit that frequency, and why it must not predict a new individual's outcome. The evaluator receives the exact aggregate of frozen states and a separate consultation prose reference; the generator receives neither reference.

Released evidence. At nominal $\epsilon = 4$, the GEPI packet supplies noisy cells including $A = 26.810150$, $C = 294.617853$, and $U = 7.885353$, together with the remaining state cells, release identity and noise scale. Tab-PE supplies its synthetic table's reconstructed cells. Aug-PE supplies five retrieved synthetic consultations with an explicit instruction that they are not a population sample and contain no aggregate counts. The public-only packet supplies no corpus measurement.

Table 4: Legal example: translated population-answer excerpts and original complete-answer scores. All four outputs answer the same request with the same GPT generator and corpus-question judge.

Method	Score	Candidate excerpt (English translation)
GEPI	5	"The proportion with a final fine disposition is roughly 5% (a very rough range of about 0–10%)." The answer explains that UNKNOWN prevents definitive classification and that the frequency does not predict a new case.
Tab-PE	3	"About $104/634 \approx 15\%$." It describes this as an approximate observation subject to UNKNOWN outcomes.
Aug-PE	3	"The frequency cannot be calculated from the supplied evidence alone." It distinguishes five retrieved synthetic documents from a population sample.
Public-only	2	"The [method] evidence contains no actual corpus measurements." The population summary is `ABSTAIN`.

Original judgment. The GEPI judgment credits the requested frequency within the rubric's five-percentage-point allowance, its UNKNOWN qualification, and the consultation component. The Tab-PE judgment says "Part 2 is materially off." The Aug-PE judgment concludes "useful consultation, but the corpus answer misses the central resolved statistic." These are the original evaluator's assessments, not a new author score.

Author inspection. The case illustrates the distinction between giving relevant advice and resolving the explicit statistical request. Abstaining when no aggregate was supplied can accurately describe the generator's evidence boundary while receiving a lower completeness score against the evaluator's corpus target. The judge's reference access must not be mistaken for information available to the generator. The 1–5 result also includes the consultation component, so the displayed excerpt alone is not a complete reconstruction of that score.

B.2 MEDICAL: PRESERVING AN IDENTIFIED INTERVAL AND ITS MEANING

Request and reference. Within complaint partition `limb_pain`, among cases with `labs=no` and `medication=yes`, the request asks for the identified interval for `imaging=yes`. Its rule is $[A/D, (A + U + E)/D]$, with unknown outcome U and unknown eligibility E. The evaluator uses an exact frozen-state interval; the generator receives only its method's evidence.

Released evidence. The GEPI packet states interval $[0.30, 0.49]$, approximate released counts $A \approx 16$, $N \approx 26$, $U \approx 0$, $E \approx 10$, $D \approx 52$, and cell noise standard deviation 5.92. It explicitly distinguishes documentary missingness from additional DP-noise uncertainty. This is the standalone GEPI release whose historical storage key was corrected from `tabpe_eps4` (Appendix D.7), not the synthetic-table comparator. The two synthetic interfaces supply their own readouts, and the public-only condition has no corpus counts.

Table 5: Medical example: original v5 scores and verbatim excerpts from stored answers. Each synthetic answer declines to attribute its supplied interval to the requested marker combination.

Method	Score	Candidate excerpt
GEPI	5	"Identified interval for imaging=yes in the specified corpus subset is [0.30, 0.49]." It separately states that DP noise was not propagated into this interval.
Tab-PE	2	"The only interval given in the evidence is 0.11–0.56 for a different statistic in the synthetic 512-row table."
Aug-PE	2	"No direct estimate can be justified from the provided evidence for the requested partition"...
Public-only	2	"Cannot be determined from the supplied evidence; no corpus counts (A, N, U, E) were provided for this subset, so the interval for imaging=yes is not identifiable."

Original judgment. The GEPI rationale states "Scope is correct for the specified corpus subset and outcome, and the uncertainty discussion correctly notes that separate DP/measurement uncertainty was not propagated." The synthetic-method judgments identify the missing requested interval as the central defect. The original score anchors and caps are given in Appendix C.1.

Author inspection. GEPI carries both the interval and its documentary interpretation into the answer. The other answers illustrate a downstream attribution problem: they mention supplied numeric intervals but decline to use them for the requested cohort. Their claims about what was supplied are candidate text, not an independent audit that their evidence packets lacked such a binding.

B.3 A CONTRASTING LEGAL ANSWER

A separate hash-selected legal parent asks about other dispositions among sexual-offense cases with prior punishment. GEPI's saved answer receives noisy counts but returns `POPULATION_SUMMARY: ABSTAIN`, citing UNKNOWN; the original score is 3. Tab-PE's answer supplies an approximate frequency and receives 5. This shows why preserving a statistic in the release is distinct from eliciting its use in every answer. The population-task examples above illustrate a mechanism, while the full-roster comparisons quantify its observed performance.

C EVALUATION PROTOCOLS AND JUDGE ASSESSMENT

The primary score measures complete-answer alignment with the request and evaluator-only references. Numerical and semantic checks are reported separately.

C.1 EXECUTED SCORING CONTRACTS

The legal judge receives the request, candidate answer, and reference, and scores the consultation and population components holistically. It permits accurate point, verbal-range, and interval statements, with five-percentage-point allowance for the population headline; no fixed additive weighting separates the two components. A historical frequency must not become an individual prediction or a causal claim. The medical judge additionally receives deterministic endpoint comparisons and a computed score cap. It must use those comparisons rather than redo the arithmetic. A documentary identification interval and an interval accounting for DP noise are different objects.

Table 6: Executed 1–5 score anchors, abridged without changing their ordering. The complete frozen instructions are indexed in Appendix I.5.

Score	Legal complete answer	Medical interval answer
5	All substantive facts and limitations correct	Supported estimate and uncertainty within 5 percentage points; correct scope; no material omission
4	Accurate useful core, minor omission or imprecision	Useful supported answer; at most 10-point endpoint defect or minor limitation defect
3	Useful correct corpus content with a material gap	Useful partial answer with a missing component, or endpoint defect at most 20 points
2	Relevant but little central resolution, or substantial factual defect	Little usable empirical answer, unjustified abstention, or larger numerical/scope defect
1	Unsupported or wrong central conclusion, irrelevant or unusable	Wrong cohort/direction, incompatible central claims, fabricated certainty or unusable answer

Medical guardrails and retention. Point-only answers or a missing requested interval are capped at 3; abstention on a nonempty row and wrong scope are capped at 2. Numeric-error bands, within-answer conflicts, and misreading a lower bound as a central estimate also constrain the supplied cap. The earlier v4 implementation additionally capped interval-only answers at 3; v5 removes that extra ceiling. The medical GPT block uses v5 and Qwen uses v4. Structurally valid raw scores above a prompted cap are retained under the disclosed post-hoc aggregation policy, rather than selectively discarded or clipped. The cap remains part of the prompt that produced the judgment; retention does not remove its possible influence (Appendix F.7).

Table 7: Conditions of the displayed utility blocks. GPT/Qwen in Table 1 names the answer generator, not a separate cross-block scoring scale.

Block	Reference and judge contract	Interpretation condition
Legal GPT	Corpus-question rubric; exact population and prose consultation references	Common 300: 245 population requests and 55 shared controls
Legal Qwen	Same corpus-question rubric and reference construction	Same 300 requests; unresolved outcomes remain in completion bounds
Medical GPT	Exact frozen-state interval; deterministic comparisons; v5	Common 300 including all $\epsilon = \infty$ conditions; raw-score retention
Medical Qwen	Exact frozen-state interval; deterministic comparisons; v4	Same 300 requests; interval-only cap; raw-score retention

Failures and denominators. A generation or judge failure is unresolved, not score 1. Means use scored requests, and unresolved requests remain in the roster for completion bounds: if s is the sum of n scored values out of N, these bounds are $[(s+N-n)/N, (s+5(N-n))/N]$. Paired contrasts use their stated common targets and clustering units; they condition on the saved releases rather than averaging over new noise draws. The hosted production judge is `gpt-5.4-mini-2026-03-17`; frozen request bodies and schemas identify each instrument exactly.

C.2 What the complementary checks measure

Table 8: Complementary checks of answer content and judge agreement.

Check	Object assessed	Reported evidence
Holistic judge	Complete requested answer	1–5 reference alignment and scored counts
Endpoint audits	Numeric spans in saved answers	Conditional endpoint error and full-roster tolerance coverage (Appendix F)
Semantic calibration	Population estimate, cohort, outcome and unknown interpretation	64 saved outputs, score-hidden annotation
Lawyer assessment	Complete legal answers under the scoring rubric	37 answers; Kendall τ_b, intervals and rater means

Table 9: Fixed 64-output semantic calibration. Spearman correlation compares the frozen population-subtask ordinal with the original 1–5 score within each block. C/D counts are concordant/discordant same-target output pairs after excluding a tie on either axis; all 16 outputs per block had a primary score.

Evaluation block	**Spearman ρ**	**Same-target C/D**
Legal, GPT-5.4-mini	0.831	14/0
Legal, Qwen3.8-27B	0.277	8/5
Medical, GPT-5.4-mini	0.752	13/0
Medical, Qwen3.8-27B	0.559	8/0

Table 10: Human-gold coherence on 37 legal answers. Kendall τ_b with case-bootstrap 95% intervals; mean score of each rater in parentheses.

Pair	**Kendall $\tau_b \pm h_{95}$**
Lawyer (2.81) vs. judge GPT-5.4-mini (3.41)	$.27 \pm 0.29$
Lawyer vs. GPT-6-Astra (3.05)	$.17 \pm 0.33$
Lawyer vs. Claude Fable 5.1 (3.89)	$.27 \pm 0.34$
Judge vs. GPT-6-Astra	$.58 \pm 0.19$
Judge vs. Claude Fable 5.1	$.67 \pm 0.16$
GPT-6-Astra vs. Claude Fable 5.1	$.58 \pm 0.20$

C.3 Primary-Score-Hidden Semantic Calibration

This diagnostic uses saved outputs from before the common-300 replacement. We fixed the first four target clusters per domain under the pre-existing SHA-256 ordering and included four release methods under both generators, for 64 outputs total. A separate model-assisted reader saw the question, candidate, and evaluator reference, but not the method, generator, primary score, primary rationale, or prior parser decision. It annotated only the requested population subtask: whether an answer asserted the requested corpus estimate, preserved cohort and outcome, handled missingness, avoided an unsupported individual prediction, and fell within a fixed five-percentage-point numeric tolerance. The original campaign score was joined only after the semantic ledger was frozen.

The diagnostic supports directional agreement in three blocks but is not a human or domain-expert validation. In the legal GPT block, numerically poor and fully compatible asserted estimates both average 2.0, so its high correlation is driven largely by separating abstention from estimation. The legal Qwen block has weak association and five discordant same-target pairs; one numerically compatible answer with an incorrect treatment of unknown outcomes still received a primary score of 4. The reader also had prior aggregate-level exposure, and answer wording can reveal release style despite administrative method-label hiding. We therefore do not use this sample for method ranking, significance claims, consultation-quality validation, or cross-block score comparison. The frozen packets, annotations, score joins, and complete tie counts are recorded in the reproducibility artifact.

C.4 Lawyer Assessment and Cross-Judge Agreement

On saved outputs from the earlier legal campaign, a practising lawyer scored 37 sampled legal answers (external roster; GEPI, Aug-PE, and no-release outputs mixed) on the same 1–5 rubric as the judge, without seeing any model score; these are the human ratings in the comparison. The same frozen judge prompt, schema, and packets were then scored by two further frontier models under a first-three mechanical gate and one attempt each. Table 10 reports Kendall τ_b (shift-invariant rank coherence) with 9,999-draw case-bootstrap 95% intervals, pooled over the sample. The production judge has a positive rank-association point estimate and is about 0.6 points more lenient; all three LLM judges agree with one another more than any agrees with the lawyer. With one rater and 37 answers this is a descriptive check, not a validation of the judge.

The sample contains 26 GEPI exact-endpoint answers, 10 Aug-PE answers and one no-release answer. All three human–LLM intervals include zero. The pooled association is descriptive, not an equivalence test or an estimate of method ranking reliability across all utility blocks. The same rubric is shared by the LLM judges.

D FIDELITY LIMITS, FORMAL PROPERTIES AND PRIVACY ACCOUNTING

The case-query bounds apply to answer distributions, independently of the scoring rubric. The remaining sections give GEPI's sensitivity, release guarantee and privacy accounts.

D.1 CASE-QUERY FIDELITY: GENERAL BOUND AND PROOF

Fix the same public question and auxiliary information on $D_1 = D_0 \cup \{d^\star\}$, where $d^\star \notin D_0$. Fix a corpus-grounded correctness rule specifying which answers are acceptable given the available evidence. For each $i \in \{0, 1\}$, choose a reference distribution P_i supported on the answers accepted by this rule on corpus D_i, and let Q_i be the answer distribution of the same private mechanism on D_i. The rule specifies acceptable answers; the choice of reference probabilities is part of the fidelity target, not an empirical frequency inferred from D_i. With a unique accepted semantic answer, P_i is a point mass. In the example of Section 2.3, P_0 puts probability one on insufficient evidence and P_1 on the finding documented in $d^\star$. All these distributions use a common semantic output space. A fixed corpus-independent semantic map of generated text is post-processing, so it retains the mechanism's DP guarantee. Define $s = \mathrm{TV}(P_0, P_1)$ and $e_i = \mathrm{TV}(P_i, Q_i)$. The quantity s measures how much the required answer changes when the protected document is removed.

Overlapping reference answers. For an (ϵ, δ)-DP complete-answer mechanism, with $\epsilon \geq 0$ and $0 \leq \delta < 1$, define

$$\kappa_{\epsilon,\delta} := \frac{e^\epsilon - 1 + 2\delta}{e^\epsilon + 1}. \tag{7}$$

Then

$$\max\{e_0, e_1\} \geq \frac{1}{2}\left[s - \kappa_{\epsilon,\delta}\right]_+. \tag{8}$$

In words, any change in the required answer that exceeds the permitted change in the private answer must appear as fidelity loss in at least one corpus.

Proof of Equation 8 and Proposition 1. For any event E, write $p = Q_1(E)$, $q = Q_0(E)$ and $a = e^\epsilon$. DP applied to E and, in the reverse direction, to its complement gives

$$p \leq aq + \delta, \qquad 1 - q \leq a(1 - p) + \delta.$$

Adding the rearranged inequalities yields $(a + 1)(p - q) \leq a - 1 + 2\delta$. Interchanging the corpora bounds the opposite difference, so $\mathrm{TV}(Q_0, Q_1) \leq \kappa_{\epsilon,\delta}$. The triangle inequality now gives

$$s \leq e_0 + \mathrm{TV}(Q_0, Q_1) + e_1 \leq e_0 + e_1 + \kappa_{\epsilon,\delta}.$$

Thus $e_0 + e_1 \geq [s - \kappa_{\epsilon,\delta}]_+$. Since $e_0 + e_1 \leq 2\max\{e_0, e_1\}$, this proves the general bound. Disjoint reference support means $s = 1$, giving $(1 - \kappa_{\epsilon,\delta})/2 = (1 - \delta)/(1 + e^\epsilon)$. □

The cap on neighboring private-answer TV is sharp given only (ϵ, δ)-DP. On a two-element output space, take $Q_1(E) = (e^\epsilon + \delta)/(e^\epsilon + 1)$ and $Q_0(E) = (1 - \delta)/(e^\epsilon + 1)$; the two inequalities used in the proof hold with equality. Reference answers concentrated respectively on E and its complement also attain Equation 1. This is an instance of DP's hypothesis-testing constraints (Dong et al., 2022), not a new privacy definition.

Which corpus loses fidelity? The two-corpus bound alone does not identify D_1 as the one that fails. If the mechanism is additionally faithful when the document is absent, $e_0 \leq \gamma$, the same proof gives

$$e_1 \geq \left[s - \kappa_{\epsilon,\delta} - \gamma\right]_+. \tag{9}$$

Alternatively, for the event E of asserting the requested fact, if $P_1(E) \geq 1 - \eta$ and $Q_0(E) \leq \beta$, the DP definition directly gives $e_1 \geq [1 - \eta - e^\epsilon\beta - \delta]_+$. These are explicit absence-fidelity or groundedness assumptions, not automatic consequences of removing a document.

Stronger bound for a Gaussian evidence release. Suppose the answer is computed from $S(D) = f(D) + Z$ and the same public inputs, where $Z \sim \mathcal{N}(0, \sigma^2 I)$, with no other protected access. For the fixed neighboring pair, let $r = \|f(D_1) - f(D_0)\|_2$. Equal-covariance Gaussian distributions have TV $2\Phi(r/(2\sigma)) - 1$, and randomized post-processing cannot increase TV. Consequently,

$$\max\{e_0, e_1\} \geq \frac{1}{2}\left[s - \left\{2\Phi\left(\frac{r}{2\sigma}\right) - 1\right\}\right]_+. \tag{10}$$

A certified global sensitivity satisfies $r \leq \Delta_2$. Since Φ is increasing, replacing r by Δ_2 gives a possibly weaker bound requiring no measurement of this particular pair. Write $\mu = \Delta_2/\sigma$: the TV expression then uses $\mu/2$, while the testing tradeoff uses the full standardized shift μ. With the fact-event assumptions above, the Gaussian testing tradeoff gives the one-sided refinement

$$e_1 \geq \left[1 - \eta - \Phi\left(\Phi^{-1}(\beta) + \mu\right)\right]_+, \qquad 0 < \beta < 1. \tag{11}$$

The event indicator composed with the answer generator is a randomized test whose size is at most β; its power is at most $\Phi(\Phi^{-1}(\beta)+r/\sigma) \leq \Phi(\Phi^{-1}(\beta)+\mu)$ because $r \leq \Delta_2$, proving this refinement (Dong et al., 2022). GEPI's existing Gaussian calibration uses $\mu = \sqrt{B}/\sigma$ (Section 4.2); no new accounting parameter is needed here. This Gaussian refinement uses the release's distributional structure, which an arbitrary (ϵ, δ) certificate alone does not establish.

Interpretation and scope. TV compares probabilities of semantic answer events, not 1–5 judge scores, token-edit distance or empirical attack AUC. The references can prescribe document-grounded disclosure versus abstention; if instead they are outputs of a non-private model, the bounds concern fidelity to that model rather than truth. We do not estimate s, β, η or γ in our experiments. The general bound applies to a complete answer using its complete DP account. The Gaussian bound also covers complete answers and repeated public post-processing of one saved synopsis; a token mechanism that reopens private records requires an account for its full transcript. Public knowledge or other records can supply the requested fact, reducing the relevant separation. Conversely, a rare population statistic can depend strongly on one document. The result therefore motivates separating evidence requirements by their dependence on protected records, without declaring every Case query impossible or guaranteeing accuracy for every population query.

D.2 Declared relational release and supported reuse

Proposition 2 (Document-level release and reuse)**.** *Fix a public workload, coordinate layout and document-local map before protected measurement. Let each canonical protected unit d contribute $x(d) \in \{0,1\}^m$ with $\|x(d)\|_1 \leq B$, and let adjacency add or remove one such unit. Release $\widetilde{H} = \sum_d x(d) + Z$ with $Z \sim \mathcal{N}(0, \sigma^2 I_m)$ over all fixed coordinates, calibrated by Equation 5. The synopsis and any answers computed solely from it and public inputs jointly satisfy document-level (ϵ, δ)-DP. Reusing that saved synopsis incurs no additional privacy loss.*

Proof. For neighboring corpora the sum changes by one contribution, and binary coordinates give $\|x(d)\|_2^2 = \sum_i x_i(d) \leq B$. Thus $\Delta_2(H) \leq \sqrt{B}$ and the calibrated Gaussian release is (ϵ, δ)-DP. Answers and adaptive requests that access protected information only through the same saved release are randomized post-processing, so their joint transcript has the same guarantee. □

Boundary of the transcript. The protected unit is the canonical document/case after the declared input construction, not every record concerning one person. Random extraction may be conditioned on data-independent randomness only when locality and the contribution bound hold for every realization. Private exact counts, source spans and validation diagnostics are not part of this public transcript. Algorithm 2's data-dependent rejection, validation or release/abort status is outside this guarantee if observable; any such output requires separate analysis. The fixed public layout and noise on zero as well as nonzero coordinates prevent support selection from becoming a separate unaccounted release.

D.3 Document-local interpretation and bounded contributions

Section 4.1 describes document-local interpretation: extraction (denoted g here) and the evaluated states $(E_w(d), P_w(d))$ as one step. The implementation splits that step across four components with disjoint inputs. The separation is not an engineering convenience: each boundary is what licenses a specific claim made elsewhere in this paper.

Table 11: Components producing a document's contribution vector. The rightmost column states what each component is structurally unable to observe or change.

Component	Input	Output	Denied access to
Contract compiler	public specification	executable contracts, family layout, coordinate offsets	protected documents
Extractor g (human or model)	one document	typed facts G_d with evidence	coordinate layout; other documents
Contract evaluator	G_d and the contract	$(E_w(d), P_w(d))$, bin	document text
Contribution encoder	evaluated states, profile	$x(d)$	revising states

The release guarantee does not depend on extraction accuracy. The hypotheses of the release guarantee in Section 4.2 are that the profile and the extraction procedure are fixed, that changing one document cannot change the extracted states of any other, and that coordinates are binary with $\|x(d)\|_1 \leq B$. Correctness of $(E_w(d), P_w(d))$ appears in none of them. An inaccurate extractor changes $H(D)$, but the contribution encoder still emits a binary vector obeying Equation 3, so $\Delta_2 \leq \sqrt{B}$ and the Gaussian privacy guarantee are unaffected. Consequently an extractor may be replaced without repeating the privacy analysis, provided it remains document-local and fixed in advance. This is a statement about the guarantee only: extraction errors are not harmless, since they corrupt the released counts and therefore the answers. The restriction carried by the hypotheses is substantive, and corpus-wide vocabulary learning or cross-document prompting violates it.

Coordinate assignment must precede measurement. The compiler fixes every offset from the public specification alone, before any protected document is read. Were assignment instead driven by observed states, the identity of an allocated coordinate would itself be a function of the data, and the sparse internal representation of $H(D)$ would leak which coordinates are nonzero. Section 4.2 requires noise over the entire fixed coordinate universe for the same reason; the encoder is denied any view of other documents so that this property is structural rather than a convention.

Structural validation is not semantic validation. The contract evaluator consumes G_d and never the underlying text. This is deliberate. An evaluator with access to the narrative could let a span-matching check stand in for a judgment about meaning, and the distinction between traceability and meaning—matching an evidence span is necessary for traceability but does not certify its meaning—would no longer be enforceable by construction. Responsibility for meaning rests with g and with the independent source-gold adapter, not with the evaluator.

Error attribution. Because the components are separable, a discrepancy in a released statistic can be localized. In a separate historical medical extraction control, the profile, contribution plan, B, noise draws, and answer policy were held identical while only g changed. One extractor validated all 219 documents and reproduced every one of the 657 source-gold states, yielding an exact histogram identical to the oracle. The other had six documents rejected to the declared unknown state by schema and evidence checks; its aggregate absolute errors, summed across the three workload queries, were 4 in A and 9 in C. These errors need not arise solely from the rejected documents. Attributing that gap to extraction rather than to aggregation or release is possible only because no component downstream of g is permitted to revise its output.

D.4 Document States and Coordinate Layouts

For a contract $w \in \mathcal{W}$ (Figure 3) and document d, let $E_w(d)$ record whether d belongs to the contract's cohort and $P_w(d)$ whether it shows the outcome, each taking a value in $\{\mathsf{T}, \mathsf{F}, \mathsf{U}\}$ (true, false, unknown):

$$E_w(d) \in \{\mathsf{T}, \mathsf{F}, \mathsf{U}\}, \qquad P_w(d) \in \{\mathsf{T}, \mathsf{F}, \mathsf{U}\} \quad \text{when } E_w(d) = \mathsf{T}. \tag{12}$$

A fixed document-local extractor supplies facts and evidence; deterministic rules assign states. Missing evidence, unresolved conditions, and invalid schemas yield U, never F. Attribution and temporal

joins stay within one document. Define

$$C_w = \sum_d \mathbf{1}\{E_w(d) = \mathsf{T}\}, \qquad A_w = \sum_d \mathbf{1}\{E_w(d) = \mathsf{T}, P_w(d) = \mathsf{T}\},$$
$$U_w = \sum_d \mathbf{1}\{E_w(d) = \mathsf{T}, P_w(d) = \mathsf{U}\}, \qquad U_{E,w} = \sum_d \mathbf{1}\{E_w(d) = \mathsf{U}\}. \tag{13}$$

Define $N_w = C_w - A_w - U_w$. Among known-eligible documents, the outcome rate for $C_w > 0$ is identified within $[A_w/C_w, (A_w + U_w)/C_w]$. Unknown eligibility is recorded separately.

Contracts with shared structure form families. A fixed document-local rule assigns each document to one public partition cell $\kappa(d)$; $\kappa(f)$ is the cell served by family f, and p ranges over these cells. The family vector $x_f(d)$ has at most b_f active binary coordinates; $\bigoplus$ concatenates these vectors:

$$x(d) = \bigoplus_f \mathbf{1}\{\kappa(d) = \kappa(f)\} x_f(d), \quad H(D) = \sum_d x(d), \quad B = \max_p \sum_{f:\kappa(f)=p} b_f. \tag{14}$$

Thus $\|x(d)\|_1 \le B$ and $\|x(d)\|_2 \le \sqrt{B}$. Families outside the assigned cell receive zero contribution. Public recipes recover each declared count from H; no passages or document-level relations are published.

The following layouts implement the family vectors and contribution bounds.

Histogram families. A contract with J_f public bins, including an explicit unknown bin, contributes one unit to its assigned bin. Thus $m_f = J_f + 2$ and $b_f = 1$. Counts, signs, ages, and elapsed or stay durations use their declared coarse bins. Binning the result does not relax the contract's original temporal inclusion condition.

Small target families. For $1 \le K_f \le 4$ targets with shared eligibility, the vector of ternary outcomes belongs to $\{\mathsf{T}, \mathsf{F}, \mathsf{U}\}^{K_f}$. We allocate one coordinate to each such state and activate one coordinate. Hence $m_f = 3^{K_f} + 2$ and $b_f = 1$. Multiple targets may be true simultaneously; this is not a mutually exclusive label encoding.

Large target families. For $K_f > 4$, we use a factorized layout. The first target, chosen in the fixed public target order, has three eligible coordinates for $\mathsf{T}, \mathsf{F}, \mathsf{U}$. Each other target has two coordinates: eligible-and-true and eligible-and-unknown. Its false state activates neither. An eligible document activates one first-target coordinate and at most one coordinate for each remaining target. Therefore

$$m_f = 2K_f + 3, \qquad b_f = K_f. \tag{15}$$

The first-target block retains the shared denominator, so the false count of any other target is recoverable as $C_f - A_{f,k} - U_{f,k}$. The layout preserves all declared target marginals, not arbitrary new conjunctions between targets.

D.5 Single-release and multiple-release accounting

GEPI releases one Gaussian vector with the contribution bound from Section 4.2. With $\mu = \Delta_2/\sigma$, its exact Gaussian privacy profile at target ϵ is

$$\delta_{\mathrm{G}}(\epsilon, \mu) = \Phi\left(\frac{\mu}{2} - \frac{\epsilon}{\mu}\right) - e^{\epsilon}\Phi\left(-\frac{\mu}{2} - \frac{\epsilon}{\mu}\right).$$

We invert this monotone condition numerically and choose σ such that $\delta_{\mathrm{G}} \le \delta$ (Balle & Wang, 2018). The persistent single-release certificate covers one saved vector and its reuse, not every experimental condition over the corpus. The executed draws and the scope of their joint accounts are listed in Appendix D.7.

When more than one vector is published, the following RDP account provides a composable bound. For independent Gaussian releases of the same query, the tighter joint Gaussian profile used in Appendix D.7 combines their precisions directly. For distributions P, Q and $\alpha > 1$, the Rényi divergence is

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_{Y\sim Q}\left[\left(\frac{dP}{dQ}(Y)\right)^{\alpha}\right].$$

A mechanism is $(\alpha, \varepsilon_\alpha)$-RDP when this divergence between its output distributions is at most ε_α for every ordered adjacent pair (Mironov, 2017). For Gaussian noise with covariance $\sigma^2 I$ and sensitivity

Δ_2, $\varepsilon_\alpha \le \alpha\Delta_2^2/(2\sigma^2)$. In GEPI, $\Delta_2^2 \le B$. For compactness, define the Gaussian RDP coefficient $c_{\mathrm{G}} = B/(2\sigma^2)$; the complete all-orders bound is $\varepsilon_\alpha \le \alpha c_{\mathrm{G}}$. This scalar is numerically equal to the corresponding zCDP parameter. The composable accountant stores the RDP curve rather than treating c_{G} as the reported privacy budget; the main single-release ledger records the same coefficient only so that any future composition cannot hide its cost.

RDP costs compose additively at each order. For releases indexed by j, a valid common-δ bound is

$$\epsilon(\delta) = \inf_{\alpha>1} \left\{ \sum_j \varepsilon_{\alpha,j} + \frac{\log(1/\delta)}{\alpha - 1} \right\}.$$

For all-orders Gaussian bounds this yields $\epsilon \le c_{\mathrm{G},tot} + 2\sqrt{c_{\mathrm{G},tot}\log(1/\delta)}$, where $c_{\mathrm{G},tot} = \sum_j c_{\mathrm{G},j}$. Formal cross-method comparisons would require a common adjacency and total (ϵ, δ) budget, including all protected selection and release steps; the baseline columns here carry nominal labels only (Appendix E).

The sum applies when records may contribute to several releases; different domain names alone do not establish disjointness. Repeated use of the same released vector adds no mechanism to this sum. By contrast, private schema adaptation, a new profile, or an updated corpus may require further accounting. Data-independent extraction randomness can be conditioned on only when the document-local contribution bound holds for every realization. Exact counts, extracted facts, validation logs, and source spans stay internal. Releasing them is not justified by the guarantee for the noisy vector.

D.6 Release ledger for the displayed results

Table 12: GEPI release objects behind the current utility and attack results. The joint scope is the listed draws, not the full development transcript.

Object	Unit, geometry and noise	Guarantee and relationship
Legal 4/8-label	Case bundle; 415 coordinates; $B = 6$; $\sigma = 2.179/1.252$	Independent draws, calibrated at $\delta = 1/3000$; joint $(11.61, 10^{-5})$
Legal 1-label	Same geometry; marginal $\sigma = 7.044$	Additional noise on the saved 4-label release; no added cost when that release is already observed
Medical 1/4/8	Consultation; 3,374 coordinates; $B = 30$; $\sigma = 20.434/5.922/3.288$	Independent standalone $(1/4/8, 10^{-5})$ releases; listed joint $(9.53, 10^{-5})$
Legal attack GEPI	Saved legal 4-label vector	Reuse of $(4, 1/3000)$ synopsis; no fresh draw for the attack

D.7 Executed Release Calibrations

The budget labels in Section 4.2 use the following geometry and noise scales. Guarantees are at $\delta = 10^{-5}$ unless stated otherwise.

Legal workload. The coordinate universe is 415 cells over 30 cohort families grouped into five offence partitions. One case contributes at most one unit to each of the six families of its own partition, so $B = 6$ and $\Delta_2(H) \le \sqrt{6}$ by (3). The $\epsilon = 4$ and $\epsilon = 8$ conditions are independent releases with $\sigma = 2.179$ and 1.252. The $\epsilon = 1$ condition adds independent $N(0, 6.698^2)$ noise to the released $\epsilon = 4$ vector (marginal $\sigma = 7.044$), so it is post-processing of that release rather than a fresh draw. The conditions were calibrated at $\epsilon = 1, 4, 8$ against the looser fixed target $\delta = 1/3000$ and are named by those calibration points, but at $\delta = 10^{-5}$ the three marginal scales solve (5) at $\epsilon = 1.33$, 5.01, 9.73; those are the claimed guarantees. Because the $\epsilon = 1$ condition is post-processing, observing all three conditions reveals no more than the $\epsilon = 4$ and $\epsilon = 8$ releases together. Those two perturb the same vector with independent Gaussian noise, so their joint release is exactly one Gaussian mechanism whose precision is the sum of the two, which solves (5) at $(\epsilon, \delta) = (11.61, 10^{-5})$. This account concerns these two independent draws, not the entire development transcript. The shallow-release ledger also records an earlier same-corpus release; changing the scope name does not reset its privacy cost.

Medical workload. In the cohort-interval workload of Table 1, the coordinate universe contains 3,374 cells. One consultation case can change at most 30 coordinates, so $B = 30$ and $\Delta_2(H) \le \sqrt{30}$. The layout has 14 partitions, each with 24 ten-cell families, plus 14 failure cells: $14 \times 24 \times 10 + 14 = 3374$. A fully known case activates two coordinates in each of six eligible families and one in each of 18 ineligible families, totaling 30. A failed case activates 24 unknown-eligibility coordinates and one failure coordinate. The encoder checks uniqueness and the cap for every case. Inputs are the 1,200 original query–answer consultations, not a concatenation of their four synthetic document variants. The three reported conditions are separate standalone releases with $\sigma = 20.434,\ 5.922,\ 3.288$, the smallest scales solving (5) at $\epsilon = 1, 4, 8$ and $\delta = 10^{-5}$, so their labels are exact. For $\epsilon = 4$, the saved vector and noise draw are unchanged. The frozen histogram, source hashes and projection code identify this object as the GEPI histogram plus Gaussian noise. Its condition key had named a baseline and was corrected after scoring; the correction changes the label, not the released values. Jointly observing the three medical releases is one Gaussian mechanism at $(9.53,\ 10^{-5})$, against a basic-composition bound of $(13,\ 3 \times 10^{-5})$; earlier releases over the same corpus are outside this account (Appendix F.7). The common-300 table uses these same release objects under both generators. Earlier Qwen diagnostics use a different $\sigma = 6.36$ vector with standalone $\epsilon = 3.69$ at $\delta = 10^{-5}$; that historical result is identified separately and is not the current table's $\epsilon = 4$ cell.

Baseline labels. The private-evolution baseline columns carry nominal calibration labels, but their public-seed artifacts are not formal matched-DP certificates. Their accountant's assumption that a record contributes at most one unit per iteration is inherited from the pinned upstream implementation and is not re-proved here.

D.8 Noise-Aware Precision Policy

The released counts $\widetilde{C}, \widetilde{A}, \widetilde{U}$ are noisy sums over $0/1$ coordinate supports. Under independent per-coordinate noise the variance of a sum is σ^2 times its support size and the covariance of two sums is σ^2 times their overlap; for recipes v, r, $\mathrm{Cov}(v^\top Z, r^\top Z) = \sigma^2 v^\top r$. Write v_X and c_{XC} for those moments. The policy solves the Fieller set for a ratio $p = X/C$,

$$\big(\widetilde{X} - p\,\widetilde{C}\big)^2 \le z^2\big(v_X - 2p\,c_{XC} + p^2 v_C\big), \tag{16}$$

and reports the hull of its lower endpoint at $X = A$ and its upper endpoint at $X = A + U$, clipped to $[0, 1]$. Solving for the ratio keeps the numerator–denominator correlation in the statement, which dividing separate numerator and denominator bounds discards; taking the hull rather than the union matters because a union of the two sets can leave a gap and then fail to cover the identification interval. With $z = 2.576$, the policy abstains when the denominator is not certified positive ($\widetilde{C} - z\sigma\sqrt{|S_C|} \le 0$), when (16) has no bounded solution, and when the resulting width exceeds 0.2; otherwise it returns `ANSWER`. This provides corpus-statistic uncertainty under the stated noise model, not population sampling inference or protection from extraction error. It differs from the direct interval of (6) in three deliberate ways: the denominator is the eligible count alone rather than $C + U_E$, confidence widening is applied, and abstention is possible. The two interfaces answer under different estimands and their numbers are not comparable.

E Baseline Specifications

The comparisons distinguish released evidence formats from the representation on which they operate.

E.1 Synthetic text: Aug-PE

Aug-PE selects and varies synthetic candidates using privatized nearest-neighbor votes (Xie et al., 2024). We use a schema-adapted proposal path; the downstream adapter retrieves only released synthetic text and supplies it to the common answer generator. Private neighbors are used only by the release mechanism, not to repair its output.

E.2 Synthetic tables: Tab-PE

Tab-PE applies private evolution to categorical tables (Tran et al., 2026), using one row per protected document and the same frozen eligibility and outcome states as GEPI. Public recipes reconstruct

Table 13: Common inputs and released evidence across methods.

Comparison	Common basis	Changed interface
GEPI / Tab-PE	Frozen states and requests	Relational synopsis / synthetic-table readout
GEPI / per-query count DP	Frozen states and requests	Shared synopsis / query-specific noisy counts
GEPI / Aug-PE	Requests and final-answer evaluation	Relational statistics / released synthetic text
Raw-context token DP	Requests and downstream contracts	Record-local token aggregation and sampled text

evidence from synthetic rows, retaining unknown states and the source-population denominator. This comparison holds document interpretation fixed.

Privacy conditions and non-private endpoints. Tab-PE/Aug-PE provide nominal-budget comparisons of synthetic evidence; the historical public-seed artifacts are not certified matched-DP releases. Appendix D.7 specifies their contribution bounds, initialization, iterations and entropy conditions. At $\epsilon = \infty$, synthesis noise is removed while source-population size and unknown states are retained; synthesis is still distinct from exact aggregation (Appendix E.4). No constraint repair consults private source values, and reusing a saved synthetic release is post-processing. The count-DP controls below instead hold document interpretation fixed and test workload sharing, including common-format and partition-aware controls.

E.3 Per-query count DP on GEPI states

Per-query count-DP ablation. For each supplied measurement, one protected record contributes to at most one of four mutually exclusive cells A, N, U, E (eligible with the outcome, eligible without it, eligible with unknown outcome, unknown eligibility); records outside the requested cohort contribute zero. Releasing these four cells has add/remove ℓ_2-sensitivity one. Each finite condition adds independent Gaussian noise, clips negative cells to zero, and forms the same identified interval used by the evaluator contract. This query-specific mechanism shares GEPI's validated record-local state representation and is therefore a representation-controlled ablation, not a repair or replacement of raw-context token DP.

Legal calibrates 332 queries (245 utility queries plus 87 reserved attack queries) at $\delta = 1/3000$, with $\sigma = 75.122, 20.275, 11.000$ for the $\epsilon = 1, 4, 8$ standalone conditions. Medical calibrates 300 queries at $\delta = 10^{-5}$ with $\sigma = 84.880, 22.449, 11.957$. Under the disclosed zCDP conversion, jointly publishing all three finite conditions gives $\epsilon = 9.408$ in legal and 9.304 in medical at their stated deltas. At $\delta = 10^{-5}$, the legal marginal guarantees are $1.193, 4.716, 9.320$ and the joint guarantee is 10.922. The 87-query reserve is accounting capacity only: the record-specific Table 2 requests have no compatible measurement mapping in this state vocabulary, so this ablation has no Table 4 attack row.

Utility comparison. At $\epsilon = 4$, per-query count noise has scale $\sigma = 20.3$ (legal) and 22.4 (medical), compared with 2.18 and 5.92 for the GEPI releases. The corresponding GEPI versus count-DP means in Table 1 are 3.322 versus 3.147 (legal GPT), 2.949 versus 2.738 (legal Qwen), 3.360 versus 2.157 (medical GPT), and 3.160 versus 2.125 (medical Qwen). Count-DP means increase strictly with ϵ in all four blocks; the legal GPT difference is only 3.386 versus 3.385 at $\epsilon = 8$. Without noise, count DP leads in three blocks, scoring 4.522 for legal GPT and 4.870 for medical GPT. This comparison supports measurement sharing under a finite workload budget. It holds the state representation fixed, not the answer-interface formatting, and does not isolate clipping relative to raw-context token DP.

Common-format and accounting controls. In medical GPT at $\epsilon = 4$, formatting the count-DP release with GEPI's evidence template gives 2.134 on 298/300 requests versus 2.157 with its own template. GEPI retains a paired advantage of 1.228 ± 0.125 points (geometry-family cluster 95% CI). Replacing global sequential accounting with parallel composition across disjoint complaint partitions lowers count-DP cell noise from 22.449 to 8.499. On one precommitted release with the same 300 requests, formatter, generator and judge, count DP scores 2.497 versus GEPI's 3.360 (paired $+0.863 \pm 0.120$; geometry-family cluster 95% CI).

Noise repetitions. In an evaluator-only 200-repetition simulation with the common interval decoder, GEPI has lower roster-mean maximum-endpoint error in every repetition (.079 versus .223 on average). At tolerance .20, mean coverage is .947 versus .537.

E.4 Non-private conditions and reuse

Removing noise does not make the information objects identical: GEPI and per-query count DP expose exact state aggregates, Tab-PE/Aug-PE retain a synthesis step, and the raw-context $\epsilon = \infty$ condition removes clipping from its token channel. Even two exact aggregate interfaces can have different packet formatting, readouts and independently generated answers; an exact ($\epsilon = \infty$) condition is not a shared universal answer ceiling.

For example, repeated requests for a supported cohort can read the same saved GEPI counts and incur no new release cost. A new conjunction absent from the public recipes cannot be recovered by declaring another request ID; it requires a new private measurement and accounting. Synthetic artifacts also permit cost-free post-processing. The distinction concerns the information available in each released object, not exclusive ownership of the post-processing property.

E.5 Reuse and Measurement Sharing

We compare shared families with separate releases at the same total ρ. Across 72 supported workload cells, joint sharing wins, ties, and loses in 30, 6, and 36, respectively, confirming that the benefit is workload-dependent. In representative high-reuse settings at $n = 100, \rho = .5$, sharing reduces standard deviation from .04047 to .01431 ($K = 2, q = 14$) and from .07794 to .03960 ($K = 4, q = 50$). Reusing a saved release incurs no additional privacy charge; a fresh release does.

E.6 Raw-context utility adapter

The utility adapter selects records by a public-similarity rule and places individual records in separate local-model contexts. It clips and aggregates transformed full-vocabulary logits with a public prior and samples up to 128 tokens for the downstream answer generator. The legal Qwen2.5-7B path reserves 245 utility plus 87 attack slots; the medical Phi-4-mini path reserves 300 utility slots. Their finite accounts give .99/3.99/7.99 at the stated domain deltas, with no refund for early EOS. Text, selected identities and logits remain internal. This is a Grislain-style adapter rather than an exact reproduction of the published backbone or implementation. The local evidence models differ from the final-answer generators named in Table 1. Even without clipping, medical GPT scores 1.899 versus 2.000 without a release. The failed population-evidence gate precludes interpreting these scores as a privacy-budget ordering; the final attack uses the different, explicitly specified channel in Appendix H.

F Additional Results and Robustness

The numerical audits use earlier frozen outputs with their stated rosters, parsers and releases. They measure retained-span error and full-roster coverage separately and do not enter the common-300 score.

F.1 Score-Blind Medical Endpoint Diagnostic

We audit a fixed sample of 30 common medical targets for each combination of generator (GPT-5.4-mini or Qwen3.8-27B) and method (public-only, GEPI, Tab-PE, or Aug-PE) at $\epsilon = 4$, for the 240 fixed outputs. Before opening method labels or evaluator references, a deterministic extractor reads the original answer fields and retains only intervals attached to an explicit identified-interval cue or unambiguous lower/upper clauses. It stores the source field, literal span, and offsets. Counts, formulas, illustrative ranges, and DP confidence intervals are not promoted to the requested interval. The resulting blind ledger is hash-frozen before the evaluator-only exact intervals are joined.

For both this audit and the legal diagnostic below, the frozen roster selects the first 30 target IDs per domain in ascending SHA-256 order of `salt|select|domain|target_id`, with salt `gepi-same-output-claim-audit-v1`. Selection uses no judge scores and requires a retained candidate in all four methods under both generators, within their common transfer pools. The medical

Table 14: Score-blind medical endpoint audit at $\epsilon = 4$. "Intervals" is the number of answers retained by the frozen candidate-interval extractor; error is the mean absolute error of the worse endpoint on those answers.

Generator	Method	Intervals / 30	Endpoint error
GPT-5.4-mini	GEPI	**20**	**0.078**
	Tab-PE	15	0.242
	Aug-PE	9	0.228
	Public-only	0	—
Qwen3.8-27B	GEPI	**30**	**0.088**
	Tab-PE	29	0.229
	Aug-PE	30	0.271
	Public-only	0	—

pool is the 240-parent Qwen transfer subset of external270, excluding its exposed pilot and the separate development addendum30. The legal pool intersects external215 with the Qwen transfer set, excluding the exposed pilot, development addendum30, and 55 consultation-only controls. Targets missing from any required method, generator, or transfer pool are excluded before hashing; thus these are common-output diagnostics, not random samples of the full 300-parent rosters.

Table 14 reports retention and endpoint error. For GPT-5.4-mini, the remaining GEPI rows comprise two no-interval and eight unresolved rows; Tab-PE has two abstention, three no-interval, and ten unresolved rows; Aug-PE has nine abstention, one no-interval, and eleven unresolved rows. Public-only has no retained span. On targets with spans retained for both methods, GEPI's endpoint error is lower on 9/10 pairs against Tab-PE and 7/7 against Aug-PE.

For Qwen3.8-27B, GEPI has lower error on 28/29 common pairs against Tab-PE and 26/30 against Aug-PE, with two ties in the latter comparison. Public-only has no retained span. Paired counts describe 30 targets, not 240 independent observations.

Retention uses all 30 rows per cell; error is conditional on a retained interval. Abstention, no-interval and unresolved statuses remain separate, with no post-hoc accuracy threshold. The parser can miss intervals split across fields or formulas and can retain a quoted interval that the answer refuses to apply to the requested cohort. The audit therefore measures numeric content, not semantic coverage, whole-answer correctness or judge validity.

An independent reproduction matched all 240 classifications, hashes, offsets, reference joins, endpoint errors, eight block summaries, and six paired comparisons. It also found one GPT Aug-PE span quoted while the answer explicitly refused to transfer it to the requested target. Removing only that false positive as a sensitivity changes GPT Aug-PE to 8/30 and 0.215 conditional error; its paired GEPI row was already unresolved, so the 7/7 comparison is unchanged. This sensitivity does not constitute a complete semantic correction.

F.2 Fixed-Output Precision–Coverage Diagnostic

We materialize a common evaluator-side ledger from retained $\epsilon = 4$ answers. A fixed regular-expression parser assigns ANSWER, ABSTAIN, PARSE FAILURE, UNRESOLVED, or GENERATION FAILURE before exact intervals are joined. Coverage is the fraction labelled ANSWER; risk is maximum-endpoint error conditional on a parsed interval. No judge score, new model call, privacy draw, protected-corpus read, or regeneration enters this replay.

This earlier diagnostic compares all three methods on the original medical 270-parent roster. At maximum-endpoint tolerance 0.20, GEPI/Tab-PE/Aug-PE retain 251/107/46 qualifying intervals, respectively, each out of all 270 parents. Rows without a parser-retained numeric interval are outside the numerator at every tolerance. The earlier legal ledger contains 245 population requests (215 originally development-separated plus 30 development-exposed). These diagnostic rosters differ from the current Table 1 rosters; the 55 consultation-only controls have no population estimand. In particular, Aug-PE's one parsed legal interval makes its conditional error uninterpretable as broad fidelity. Parser misses and semantically misassigned numeric spans remain possible, so this diagnostic is not full-answer correctness or a validation of the holistic judge. Source commitments and the complete answer-state ledger are recorded in the reproducibility artifact. The full-roster parser

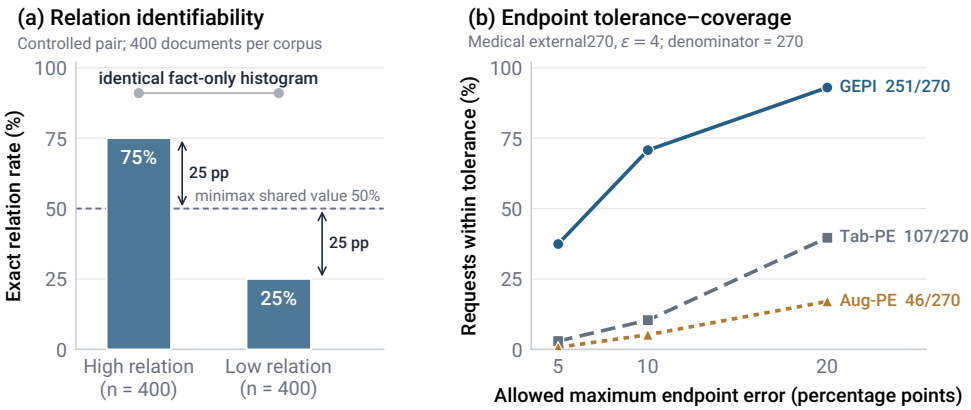


Figure 5: Relation preservation and fixed-output endpoint tolerance–coverage. **(a)** In the controlled 400-document corpus pair, identical fact-only histograms coexist with exact relation rates of 0.75 and 0.25; any shared output therefore incurs at least 0.25 worst-case absolute error. **(b)** On the common GPT-5.4-mini medical 270-parent roster at $\epsilon = 4$, each point is the fraction of all 270 requests for which the frozen parser retains an interval whose maximum endpoint error is at most the stated tolerance. Rows without a parser-retained numeric interval lie outside every tolerance. Panel (b) is a surface endpoint diagnostic, not semantic cohort accuracy or a tuned precision-policy curve.

Table 15: Fixed-parser precision–coverage replay at $\epsilon = 4$. "Ans." is literal-interval parser coverage on the fixed roster; n_r and MAE are conditional on a parsed interval with a numeric target. The last column is the fraction of those intervals with maximum-endpoint error above 0.10.

Roster	**Method**	**Ans.**	n_r	**MAE**	$\Pr(e > .10)$
Medical external270	GEPI	267/270	267	0.084	0.285
	Tab-PE	244/270	244	0.253	0.885
	Aug-PE	250/270	250	0.309	0.944
Legal population245 (diagnostic)	GEPI	224/245	214	0.067	0.121
	Tab-PE	238/245	228	0.189	0.715
	Aug-PE	1/245	1	0.027	0.000

differs from the cue-restricted 30-target extractor in Table 14; their retention counts are not directly comparable.

F.3 Evidence Fidelity and Relation Preservation

A controlled paired construction keeps relation-stripped summaries identical while changing the underlying relation prevalence. A decoder of those summaries must return the same answer for populations requiring different answers. A corresponding medical comparison removes subject and temporal conditions from the declared measurement.

In the controlled pair, 400 documents per corpus have identical fact-only signatures, while the exact same-subject diagnosis–treatment rates are 0.75 and 0.25. Any decoder of the stripped representation gives the same output for both, with minimax absolute rate error at least 0.25. This loss occurs before privacy noise. Across 512 fixed-seed Gaussian simulations, the relation-aware rate error on answered trials is 0.0008–0.0045 over the tested noise grid, so noise is not the source of the 0.50 relation gap in this control.

In the 219-document public medical source check, three fixed diagnosis–treatment measurements yield 69 gold-eligible document–measurement pairs. Eligibility specifies the population to evaluate, not a positive diagnosis–treatment relation: 12 of these pairs are gold TRUE and 57 are gold FALSE. Ignoring subject and time relations predicts 18 TRUE, including six false positives and no false negatives, and is therefore correct on 63/69 eligible pairs (91.3%); for the pneumonia and analgesic example it creates two false positives among 32 eligible pairs. The legal component study exposes a different failure: systematic interpretation errors can bias an aggregate even when document-level

Table 16: External derived LegalM utility: mean (scored n), with fixed denominator 215 per condition. Each cell uses one fixed release and one generator/judge run. Tab-PE and Aug-PE are public-seed, nominal-ϵ artifacts, not validated formal DP releases. Historical GEPI $\epsilon = 1$ is a Gaussian post-processing of the public $\epsilon = 4$ release: its marginal guarantee is $(1, 1/3000)$, while their joint guarantee remains $(4, 1/3000)$. The fresh $\epsilon = 8$ release gives the displayed subset a basic-composition bound of $(12, 2/3000)$; this is not full-transcript accounting.

Method	$\epsilon = 1$	$\epsilon = 4$	$\epsilon = 8$
GEPI	2.9442 (215)	3.1589 (214)	3.0884 (215)
Tab-PE	2.8920 (213)	2.9057 (212)	2.8122 (213)
Aug-PE	2.8140 (215)	2.7535 (215)	2.8233 (215)

Table 17: Common-scored-parent paired differences on external215. Conservative half-widths summarize descriptive 95% percentile intervals from 50,000 parent resamples (NumPy PCG64; base seed 20260924 plus a fixed pair-name hash). Unlike the cohort-cluster analysis of this historical $\epsilon = 4$ comparison, this table resamples consultation parents. It does not incorporate release-seed, generation, or judge uncertainty; no multiplicity adjustment is applied.

Difference	Paired n	Mean $\Delta \pm h_{95}$
GEPI1 − Tab-PE1	213	$+0.0376 \pm 0.1784$
GEPI1 − Aug-PE1	215	$+0.1302 \pm 0.1349$
GEPI4 − Tab-PE4	211	$+0.2559 \pm 0.1706$
GEPI4 − Aug-PE4	214	$+0.4065 \pm 0.1496$
GEPI8 − Tab-PE8	213	$+0.2864 \pm 0.1784$
GEPI8 − Aug-PE8	215	$+0.2651 \pm 0.1349$
GEPI8 − GEPI4	214	-0.0654 ± 0.1775
GEPI4 − GEPI1	214	$+0.2196 \pm 0.1542$
GEPI8 − GEPI1	215	$+0.1442 \pm 0.1582$

agreement appears acceptable. These observations support separate validation of each stage; they do not establish general graph-extraction accuracy.

F.4 Historical External Legal Corpus-QA Budget Comparison

This earlier campaign retains 215 consultation parents, corpus-frequency questions, measurement bindings, and holistic 1–5 judge for the $\epsilon = 1, 8$ successor to the $\epsilon = 4$ evaluation. Independent reaggregation checks question identity across all ten conditions, the new committed candidates and raw judgments, and the existing $\epsilon = 4$ statistics. Table 16 reports scored-only means with the original denominator of 215 in every condition. The no-release control is 2.8744 (215/215 scored).

There are four unresolved $\epsilon = 4$ rows (one GEPI, three Tab-PE), plus two Tab-PE rows at each of $\epsilon = 1, 8$. The four new failures are provider output-token truncations: one generation and one judge failure per Tab-PE condition. Assigning unresolved rows any score in $[1, 5]$ gives all-215 mean bounds of $[3.1488, 3.1674]$ for GEPI4 and $[2.8744, 2.9116]$, $[2.8791, 2.9349]$, and $[2.7953, 2.8326]$ for Tab-PE1/4/8. All other means are fully observed. These missing-outcome bounds are distinct from the sampling intervals below.

A complementary historical $\epsilon = 4$ comparison resamples the 53 cohort IDs with replacement, retaining all parents in each selected cohort (50,000 replicates). On the same common-scored parents, GEPI4 minus Tab-PE4 is $+0.2559 \pm 0.1542$ (cohort-cluster 95% CI) ($n = 211$), and GEPI4 minus Aug-PE4 is $+0.4065 \pm 0.1656$ ($n = 214$). The respective Python random-generator seeds are 4004076916 and 560564601. These intervals account for dependence within the declared cohorts but still condition on one realized release, generation, and judge run per condition. The parent-bootstrap intervals below use a different resampling unit and are not replacements for these intervals.

The observed GEPI curve is nonmonotone: $\epsilon = 8$ has 58 wins, 90 ties, and 66 losses against $\epsilon = 4$ on their 214 common scored parents. The paired interval includes zero, so these runs establish neither a monotone trend nor a statistically resolved reversal. The $\epsilon = 8$ comparisons against both baseline artifacts have positive intervals; the two $\epsilon = 1$ intervals include zero.

F.5 SCORE-BLIND LEGAL NUMERIC-SURFACE DIAGNOSTIC

We apply an analogous API-free diagnostic to 30 common legal population targets for each generator and method at $\epsilon = 4$ (240 fixed outputs). Before opening the evaluator key, the procedure writes the literal Part 2 answer state, population-summary field, missingness field, and offsets in a separate step. It then compares an unambiguous point or two-endpoint range with the evaluator's identified set using the predeclared five-percentage-point allowance. The parser excludes verbal, abstaining, and malformed surfaces and attempts to exclude qualified alternatives; judge scores, caps, verdicts, and rationales are never read.

For GPT-5.4-mini, GEPI and Tab-PE both expose a numeric surface on 30/30 targets, with 28/30 assessable against an identified reference. GEPI is within the allowance on 23/28 and has conditional mean distance 0.038, compared with 6/28 and 0.139 for Tab-PE. GEPI has lower distance on 27/28 common assessable targets. For Qwen3.8-27B, GEPI exposes a numeric surface on 26/30 targets (23 assessable; 19 within allowance; mean distance 0.044), while Tab-PE does so on 29/30 (12 assessable; 2 within allowance; mean distance 0.166). GEPI has lower distance on all 10 common assessable targets. Public-only and Aug-PE mostly abstain, so no stable paired numeric comparison to those conditions is available. The sole GPT public-only numeric row is a public-prior alternative rather than a corpus-grounded estimate; excluding that independently identified false positive leaves no comparable public-only row and does not change either GEPI–Tab-PE comparison.

These are conditional literal-field diagnostics, not whole-answer accuracy or an independent validation of the 1–5 judge. In particular, they do not establish that a surface number preserves every cohort and outcome condition, nor do they score useful verbal or qualified answers. Eighteen Qwen rows with interval expressions are parser-unresolved because they repeat equivalent decimal, percentage, or fractional forms (3 GEPI and 15 Tab-PE), making the conditional Qwen comparison wording-selected. Four numerically compatible GEPI examples also explain UNKNOWN outcomes incorrectly; endpoint compatibility therefore does not validate missingness reasoning. GPT uses a direct population request, whereas Qwen combines consultation and population sections, so the two blocks are not a pure generator comparison. The current hashes and all joins were independently reproduced, but the join code does not enforce a pre-key frozen- ledger hash. The ledger, per-row exclusions, and independent audit are retained with the diagnostic artifact.

F.6 COMMON-300 COMPARISONS AND MISSING OUTCOMES

The final bundle contains four blocks, 84 cells and 25,200 request–condition rows. Table 18 records every scored count and all-300 completion bound. Means and confidence intervals use scored rows; completion bounds preserve all unresolved generation and judge outcomes. Legal clustering uses the shared population statistic, with consultation-only controls as singleton clusters; medical clustering uses the geometry family. The saved aggregate also reports the predeclared coarser legal cohort and medical complaint-partition sensitivity, paired contrasts, retained/new strata and execution waves. Intervals come from a 10,000-draw cluster bootstrap conditional on the saved releases, without resampling release noise. We report each as the mean $\pm$ the larger distance to either end of its 95% percentile interval.

At legal GPT $\epsilon = 4$, the paired GEPI gains are $+0.704 \pm 0.12$ over Tab-PE, $+0.352 \pm 0.11$ over Aug-PE, and $+0.322 \pm 0.11$ over public-only generation (297/298/298 scored pairs). Clustering more coarsely by cohort gives $+0.704 \pm 0.14$ and $+0.352 \pm 0.14$ for the first two, preserving the primary comparison's direction. Assigning unresolved outcomes their worst-case values leaves all-300 gains of at least $+0.670$, $+0.333$, and $+0.303$ over Tab-PE, Aug-PE and public-only, respectively. The 55 consultation-only controls share outputs and contribute no between-method difference. On the predeclared newly added medical stratum (60 requests per generator), GEPI also has the highest scored mean among all methods at each finite budget, $\epsilon \in \{1, 4, 8\}$. This descriptive check separates the new requests from the retained 240; it uses the same corpus and measurement families. Worst-case completion bounds preserve the ordering in five of the six comparisons; they overlap for Qwen at $\epsilon = 1$. The reported best mean is not automatically a statistically separated or missingness-robust winner. For example, legal GPT at $\epsilon = 8$ has GEPI 3.386 and per-query count DP 3.385, whose completion bounds overlap. The largest unresolved count is 53/300 for medical Qwen per-query count DP at $\epsilon = 8$; legal GPT's exact GEPI endpoint has 33/300 unresolved. These outcomes remain visible in the complete count table rather than being converted into scores.

Table 18: Scored requests n/N and fixed-denominator completion bounds (unscored → 1 and 5) for every cell in Table 1.

Method	$\epsilon = 0$	$\epsilon = 1$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = \infty$
		Legal — GPT-5.4-mini generator			
Public-only	299/300 [2.99,3.00]				
GEPI		297/300 [3.10,3.14]	298/300 [3.31,3.33]	295/300 [3.35,3.41]	267/300 [3.65,4.09]
Tab-PE		299/300 [2.60,2.61]	298/300 [2.61,2.64]	299/300 [2.68,2.69]	299/300 [3.96,3.97]
Aug-PE		299/300 [3.02,3.04]	299/300 [2.96,2.97]	299/300 [2.98,2.99]	299/300 [2.98,2.99]
Per-query count DP		299/300 [2.78,2.80]	299/300 [3.14,3.15]	299/300 [3.38,3.39]	299/300 [4.51,4.52]
Raw-context token DP		299/300 [3.10,3.11]	298/300 [3.02,3.04]	299/300 [3.08,3.09]	299/300 [3.08,3.09]
		Legal — Qwen3.8-27B generator			
Public-only	298/300 [2.65,2.68]				
GEPI		299/300 [3.05,3.07]	295/300 [2.92,2.98]	299/300 [3.12,3.13]	296/300 [3.87,3.92]
Tab-PE		296/300 [2.14,2.19]	297/300 [2.20,2.24]	297/300 [2.23,2.27]	295/300 [3.79,3.86]
Aug-PE		296/300 [2.63,2.69]	296/300 [2.59,2.65]	297/300 [2.63,2.67]	297/300 [2.72,2.76]
Per-query count DP		294/300 [2.53,2.61]	298/300 [2.73,2.75]	296/300 [2.95,3.00]	299/300 [3.89,3.90]
Raw-context token DP		298/300 [2.74,2.76]	299/300 [2.72,2.73]	295/300 [2.67,2.74]	299/300 [2.67,2.69]
		Medical — GPT-5.4-mini generator			
Public-only	300/300 [2.00,2.00]				
GEPI		300/300 [2.66,2.66]	300/300 [3.36,3.36]	296/300 [3.63,3.68]	300/300 [4.34,4.34]
Tab-PE		300/300 [2.25,2.25]	300/300 [2.43,2.43]	300/300 [2.46,2.46]	296/300 [2.34,2.40]
Aug-PE		300/300 [2.08,2.08]	300/300 [2.16,2.16]	300/300 [2.15,2.15]	300/300 [2.17,2.17]
Per-query count DP		300/300 [1.93,1.93]	300/300 [2.16,2.16]	300/300 [2.47,2.47]	300/300 [4.87,4.87]
Raw-context token DP		298/300 [1.97,1.99]	295/300 [1.94,2.01]	300/300 [1.96,1.96]	298/300 [1.89,1.92]
		Medical — Qwen3.8-27B generator			
Public-only	300/300 [2.00,2.00]				
GEPI		299/300 [2.56,2.57]	300/300 [3.16,3.16]	299/300 [3.20,3.22]	300/300 [3.56,3.56]
Tab-PE		300/300 [2.14,2.14]	296/300 [2.26,2.31]	299/300 [2.30,2.31]	299/300 [2.32,2.34]
Aug-PE		282/300 [1.90,2.14]	286/300 [1.96,2.15]	281/300 [1.95,2.20]	278/300 [2.03,2.33]
Per-query count DP		273/300 [1.86,2.22]	280/300 [2.05,2.32]	247/300 [2.10,2.80]	288/300 [3.51,3.67]
Raw-context token DP		300/300 [1.97,1.97]	300/300 [1.94,1.94]	293/300 [1.88,1.97]	289/300 [1.87,2.02]

F.7 Medical Scoring Policy and Historical Instruments

The medical GPT contract removes the earlier v4 cap of three on interval-only answers, because the task explicitly requests an identified interval. The v5 successor passed a 15-control calibration; Qwen retains its frozen v4 contract. Both blocks retain structurally valid raw scores above a prompted cap under the disclosed aggregation rule. This removes selective exclusion of valid scores but does not undo the cap's influence on the original judgment. All current methods and endpoints use their block's same 300 requests.

Separate semantic-score diagnostic. An earlier semantic-v5/score-v6 instrument used a distinct semantic parser and score contract on external270 plus 30 development-exposed requests. Its numeric rubric-band prediction agreed with the judge on 971/1007 scored external270 answers. That instrument's means and scale-check extensions are preserved in the historical artifact; they are not validations or replacement scores for the current interval-cap-v4/v5 blocks.

The historical interval-cap campaign also compared two GEPI releases, one then stored under a Tab-PE key. That comparison is not evidence about the synthetic-table baseline. Its earlier finite-only transcript had a conservative basic-composition bound of $(26.28678, 6\times10^{-5})$; this is a historical accounting scope, not the joint guarantee for the displayed release ledger. Any transcript including the non-private exact evaluator channel has no finite-DP guarantee. Complete earlier tables and instrument revisions remain in the archive, separately from the common-300 result.

G Evidence-Requirement and Preservation Audits

G.1 Taxonomy examples

The classes describe evidence requirements independently of access rights, not mutually exclusive request templates. An internal leave rule remains K; checking the contents of a published case record remains Case; and an empirical adverse effect relationship remains P even when described in a public source. In a legal setting, "Is settlement generally mitigating?" is K; "Among first-time defendants who settled, what fraction received custody?" is P; adding "I am a first-time defendant who settled" makes the population comparison CP; and "What sentence and rationale appear in my judgment?" is Case. "Explain settlement, compare me with similar cases, and quote my judgment"

Table 19: Semantic evidence needs, defined independently of source access. Any type may retrieve from a corpus treated as protected; access policy separately governs use of each source or release.

Type	Answer target	Information granularity
Knowledge (K)	Fact, definition, explicit rule, or rule application; not an empirical population relation	Proposition and its scope, conditions, exceptions, and authority
Population (P)	Empirical frequency, pattern, risk, or association across records, possibly asked qualitatively	Linked units, conditions, outcomes, denominators, and missingness
Case–population (CP)	An individual's or case's conditions used to interpret empirical population evidence	A matched population relation and the conditions used to match it
Case	The contents, outcome, reasons, or attribution of a designated record, including a patient file or specific contract	Attributed record detail and supporting passages

is a compound K+CP+Case request. CP requires that the case attributes define or interpret the comparison cohort; adjacency between unrelated Case and P questions is insufficient. Personalizing a general explanation with only user-provided context remains K. "Which rule applies to my stated facts?" is K absent a request for empirical outcomes. "How do cases like mine usually end?" is CP even without an explicit percentage or named comparison group. These component labels are used to analyze the request. The main GEPI answer path receives the original request and does not run a K/P/CP/Case classifier.

Rubric migration diagnostic. An earlier rubric labelled all 20 medical development requests K-only. Under the revised rubric, two passes over the same 20 requests labelled all 20 CP; the requests ask for an empirical interpretation of the user's conditions even when no numerical statistic is requested. This is a *non-blind development migration*, using the same model family and previously reviewed items, not an independent estimate of CP prevalence in held-out questions. The unchanged reference answers remain mainly clinical advice: only one of 20 contains a population-statistic expression, and it is not a frequency claim. Hence the semantic relabelling alone neither makes the existing gold answers reward a statistical response nor establishes that the corpus records the required outcomes. On eight predefined rubric anchors, the two stances reproduced seven and eight labels, respectively; both retained the three clear K anchors. The candidate medical main workload was subsequently excluded by its pre-execution family and reference-quality gates, so this migration was not promoted to a main result. In calibration, we distinguish asking *what rate a named guideline reports* (K: specified source content) from asking for the *actual rate* in a population (P); one of two development stances missed this boundary.

Prior private-RAG evaluation: synthetic medical reconstruction. The inspected pre-publication code of Grislain (2025) asks for a disease label from a speaker's symptom list. We reconstructed 200 sampled source rows from a pinned dataset, but they contain only 63 distinct question strings; repeated strings are not independent formulations. An OpenAI-family annotation that grouped seven questions per call showed a batch effect: whole batches received one label although the question template was fixed. A one-question-per-call variant was non-blind, because the earlier results were already visible to the procedure that issued it, and it still assigned different classes to the same template. We therefore exclude both OpenAI label proportions.

A separately prompted Anthropic-family model, given no gold answers or protected passages, proposed CP for all 63 strings (200 rows). Two strings (10 rows) had been inspected beforehand, leaving 61 strings (190 rows) actor-blind. Five strings (21 rows) were flagged ambiguous because some synthetic symptoms are implausible in real medicine. A separate *non-blind, model-only* task-template adjudication classified the request as CP without semantic ambiguity: the speaker supplies conditions to apply a corpus-derived symptom–disease relation, and no designated case record is being checked. Implausible symptoms affect real-world answerability, not the type of evidence requested in this synthetic task. This is not human ground truth or an inter-model reliability estimate.

A post hoc sample expansion enumerated all 97 distinct question strings in the reconstructed 5,100-row evaluation prefix. We carried forward the 63 identical-string labels and applied the same one-question-per-call instrument to the 34 new strings. All 97 received CP proposals (six flagged am-

Table 20: Corpora and question units used in Section 2. "Reconstructed" denotes questions recovered from an earlier study's released materials without verifying its exact original evaluation run. "New sample" denotes our own question selection and corpus boundary, not a prior paper run. The first five rows use benchmark families evaluated in prior private-RAG work; ContractNLI is an additional Case control from a legal-RAG benchmark. The benchmark corpora play a protected-source role in the audit; we do not claim that the files themselves are confidential.

Evaluation questions	**Corpus supplying source evidence**	**Questions and recovery limit**
DPVoteRAG NQ	Retrieved Wikipedia passages	100 reconstructed questions; original corpus snapshot unverified
DPVoteRAG TriviaQA	Retrieved Wikipedia passages	100 reconstructed questions; original corpus snapshot unverified
Grislain synthetic medical	Synthetic symptom–disease *doctor-response* documents	97 distinct questions in the reconstructed 5,100-row run; paper-run size differs from released code
SearchQA (prior benchmark, new sample)	24,914 SearchQA validation-set search snippets with document hashes	408 selected taxonomy questions; all 408 audited on the web across two stages
ChatDoctor (prior benchmark, new sample)	119,078 remaining HealthCareMagic/iCliniq patient Q&A pairs with document hashes	408 selected questions; original reported split unavailable
ContractNLI (Case control, new sample)	95 content-hashed LegalBench-RAG nondisclosure agreements	408 selected questions across agreements; no prior private-RAG run
Earlier ClinSQL	Structured MIMIC-IV clinical tables (SQL task)	190 earlier taxonomy items; 340 separate preservation items, two audit strata
Earlier Korean LegalQA	Indexed collection of 1,830 Korean legal Q&A pairs	200 earlier taxonomy items; 400 separate preservation items, two audit strata

biguous). The 97 strings still share one symptom-to-diagnosis template and are not 97 independent query types. The author code and pinned data support a reconstruction of the code-run questions, but the original paper-run item log and protected-corpus boundary are unverified. These labels neither show that independent public evidence is available nor that GEPI's declared statistics could answer the questions.

Study-setting inventory and its denominator. The Query–Access audit includes reconstructions of earlier private-RAG evaluation settings and newly selected samples. In the reconstructed settings, we keep each question paired with its protected-corpus role. Table 20 distinguishes the two sources of questions; NQ and TriviaQA are separate settings of the same DPVoteRAG paper, not independent papers. The author retrieval script uses a RAG Wikipedia index and writes per-question passage files, but neither those retrieved files nor a content-hashed index revision was released in the inspected commit. Their protected corpus boundary is therefore unverified. Grislain's code indexes 10,000 synthetic doctor-response documents whereas the paper describes roughly 5,000; the selected 200 rows reconstruct a code run, not the exact paper run. Public benchmark availability does not change the corpus's *protected role* in these evaluations.

New content-hashed proxy settings. Because the original SearchQA and ChatDoctor evaluation IDs and full corpus boundaries could not be recovered from the inspected studies, we built two new settings from released datasets and added a third to probe designated-record legal questions. These are our own content-hashed samples, never private-RAG paper-run reconstructions. For taxonomy, SearchQA uses hash ranks 102–509 of the pinned `kyunghyuncho/search_qa` validation archive, disjoint from a preceding 102-question development sample. The 408 questions link to 24,914 nonempty snippets with document IDs and content hashes. Its question includes the Jeopardy category; leaving that category out would remove constraints from some clues. Answer-free model passes under the same rubric proposed K for 407/408 questions and P for one demographic-comparison clue; six were marked semantically ambiguous. The original and added 204-question strata contained 204 and 203 K proposals, respectively. The current-web answerability audit covers these 408 questions across two audit stages, as described below.

SearchQA public-evidence audit. Across two audit stages, public evidence yields 351 benchmark-aligned answers out of 408 requests (86.0%). A separate reference-answer-blind judge rates 272 of these fully supported by their citations (66.7% of all 408), 77 partially supported, and two unsupported. These are descriptive pooled counts, not an estimate from a single uniform search protocol.

A searcher receives only the category and clue, never the reference answer, protected-URL index, or exclusion list. Validator-side screens remove protected URLs, pages matching normalized 90-

character probes from protected snippets, and apparent benchmark derivatives. The derivative screen combines literal 60-character clue windows, local clue-token overlap, at least two quiz-show markers on a page containing the answer, and category–answer proximity. Gold-dependent filters only remove pages; the gold answer is never supplied to the searcher or reader. These screens reduce contamination risk without establishing complete source independence. A reader answers from admitted excerpts or abstains; each quotation must match the excerpt actually supplied. One judge checks benchmark equivalence; a separate judge sees the clue, candidate answer, and quotations but no reference answer and checks whether those quotations support the answer.

The first stage covered 102 requests. Retrieval and quotation handling were refined after inspecting failures, and selected requests received a second search pass. The remaining 306 requests were evaluated with the extension roster, prompts, models, and analysis hashed before observing its outcomes, without further outcome-selected searches. Each extension request received one search request with at most two web-search calls and 20 candidate URLs. Exact protected URLs and redirects were excluded against the full 408-question boundary. The first eight successfully fetched pages per question were screened, without replacing flagged pages. Of 2,395 page occurrences within that budget, 267 were flagged and 2,128 admitted; six questions had no admitted page. Readers received up to three excerpt windows per page. Search and reading used `gpt-5.5-2026-04-23`; alignment and support judges used `gpt-5.4-mini-2026-03-17`. The reader output cap was 1,800 tokens in the extension versus 900 in the earlier stage. Each extension stage checked its first three requests for HTTP success, model identity, schema validity, and finite usage before dispatching the rest; these gates did not select on answer content or scores.

All 408 requests remain in the denominator: 367 produced answers with valid quotations, of which 16 disagreed with the benchmark; 37 were abstentions, and four failed quotation validation. One extension search response was incomplete and retained without retry. Neither failures nor abstentions were removed from the reported rates. The extension had already been classified for taxonomy, so it is not a wholly unseen sample. The pooled rates characterize these questions and the retrieved current-web evidence, not knowledge requests in general or historical web availability.

For ChatDoctor, we extended the original salted-hash selection to 408 questions, fixing each added 204-question stratum before its taxonomy calls, and hash-registered the remaining 119,078 HealthCareMagic and iCliniq patient Q&A documents as the protected corpus role. The inspected papers describe a different, roughly 200,000-item corpus and do not identify their 100 evaluation questions, so this setting cannot be used to reproduce their result. With the same taxonomy instrument, the combined component sets were CP in 322 requests, CP+K in 40, CP+K+P in three, CP+P in seven, CP+Case in one, K in 30, P in two, K+P in two, and Case+K in one. Thus 373/408 contained CP, and 44 requests were flagged ambiguous. The first and added strata contained CP in 191/204 and 182/204, respectively, so the combined fraction is reported as a descriptive expansion, not an iid precision gain. These are proposed model labels, with no human adjudication or evidence that the corpus supports the requested clinical inference. We did not run a public-source sufficiency search on these patient-specific consultation requests; the public-web SearchQA audit instead covers all 408 SearchQA questions across two stages. Both settings are our own proxies and are kept separate from the registry of published DP-RAG runs.

For the legal Case contrast, we used the official LegalBench-RAG ContractNLI archive (Pipitone & Houir Alami, 2024). It contains 95 nondisclosure agreements and 977 questions with source-document character spans; all 1,389 provided spans matched the archived documents exactly in our check. Because the archive has no question IDs, a question's zero-based position in its `tests` array is its source ID. We first selected the lowest salted-hash question for each agreement, then the 313 lowest-hash remaining questions, fixing 408 questions across all 95 documents. The original 204 selections remain unchanged. Document IDs and content hashes define the protected-corpus role. One answer-free model pass with the same rubric proposed Case for all 408, with no ambiguity flags. Questions from the same agreement are clustered by document, so 408 questions do not represent 408 independent NDA documents. The questions ask what a specified NDA indicates; they share a question template, so this unanimity is not an inter-annotator reliability result. We did not test public evidence sufficiency or infer that a DP statistic can reproduce NDA wording. The official archive and our selected proxy are distinct from any prior private-RAG evaluation run.

The wider source registry also records settings not yet eligible for an item-level denominator: DPVoteRAG ChatDoctor (questions unreconstructed), DP-KSA's Wikipedia-backed NQ/TriviaQA

Table 21: Exploratory answer-string evidence in current non-Wikipedia pages for reconstructed DPVoteRAG K questions. "Answer-stating" is a model judgment with an exact page span and compatible reference-answer string, *not* a verified FULL public-evidence verdict under the frozen rubric.

Setting	K proposals	Eligible	Answer-stating
NQ	20	19	11
TriviaQA	20	18	12
Total	40	37	23

(item IDs unavailable), DP-SynRAG's synthetic medical, MovieLens, and SearchQA tasks (evaluation manifests unverified), and MuRAG's NQ/TriviaQA, MQuAKE sequences, and ChatDoctor (original run IDs unavailable). We neither replace these with example questions nor pool repeated NQ/TriviaQA source families as independent settings. The registry records source URLs, versions, task boundaries, and recovery grade. The old ClinSQL and Korean LegalQA samples are separate evidence-preservation workloads, not members of the new prior-DP-RAG query denominator.

Under the earlier, superseded single-label rubric, 190 ClinSQL taxonomy items were assigned P for 3 and CP for 187, while 200 Korean LegalQA items were assigned K for all 200. These figures describe the old rubric and must not be pooled with or interpreted as v2 component labels. The separate expanded preservation rosters contain 340 ClinSQL and 400 Korean LegalQA items, respectively, in two audit strata per setting; 308 and 327 are assessable after excluding zero-required rows and terminal schema failures (Table 23).

Current public-source K pilot. We sampled 20 NQ and 20 TriviaQA questions from the reconstructed DPVoteRAG candidate sets, excluding the first 12 hash-ranked items per setting that had been exposed during exploratory search. The answer-free taxonomy view went to two model-family passes before the public-source judgments; both proposed K for all 40. One pass marked three cases semantically ambiguous, so the conservative pure-K subset contains 37 questions (NQ 19, TriviaQA 18). These are proposed semantic labels, without a human label audit. The same Anthropic-family rubric prompt proposed CP for all 97 distinct Grislain medical questions; this contrasts settings but does not establish agreement reliability. The two model families both proposed K on the same 40 questions, but this is retrospective cross-family concordance, not contract-compliant inter-model agreement. The 24 exploratory-exposure exclusions and 40 later audited questions now have a *retrospective* exposure ledger, which cannot replace a manifest frozen before calls. The reconstructed NQ/TriviaQA task contracts were also not frozen before these passes. Thus these labels remain development proposals, not contract-compliant v2 annotations.

For each original question, a searcher generated candidate current-web URLs without reference answers. A separate validator saw the reference answer aliases, retrieved page text, and the question, then judged whether the pages appeared to state the answer. We excluded the protected Wikipedia/Wikimedia/Wikidata family and recognizable benchmark derivatives before source admission. A full verdict required a compatible answer and an exact cited page span; code checked both the span's presence and its answer alias. However, the searcher's prompt also named forbidden-source families, contrary to the frozen separation between searcher and validator; source selection was consequently biased by that knowledge. Of 181 candidate URLs, one was excluded, 60 beyond the top three per question were not fetched, 39 fetch attempts failed, and 81 pages were admitted. Two questions had no admitted page. A deterministic alias-string check found an answer alias somewhere in an admitted page for 36/40 questions, but this does not establish that the page expressed the queried relation.

The pinned model validator called 23/40 excerpts answer-stating, 4/40 partial, and 13/40 unresolved. All three semantically ambiguous items were unresolved, leaving 23/37 answer-stating, 4/37 partial, and 10/37 unresolved in the conservative pure-K subset; NQ contributes 11/19 and TriviaQA 12/18. The validator was not seeded: three independent runs on the same 40 items gave 21, 24, and 23 answer-stating judgments. An OpenAI-family validator disagreed on 13/40 items, and model-only review of those disagreements raised the apparent count to 26/37 (27/40 overall). The two validators used different implicit thresholds for scope, attribution, and presuppositions because required answer atoms were not frozen before search. These counts are exploratory answer-locatability diagnostics, not whole-request evidence sufficiency, generated-answer accuracy, or human adjudication.

A question-only reinspection enumerated 65 required answer atoms across the 40 questions: 20 requests have one atom, 16 have two, three have three, and one has four. For example, the *Rite of Spring* TriviaQA item combines three events and asks for their common year; a page stating one year or one event does not answer the full request. We also found three reference answers with the wrong answer type and four damaged question strings. A gold-alias hit can therefore both miss a genuinely relevant source and reward an irrelevant one. Both the atom list and the record of gold-answer defects were compiled *after* the original search, so neither can retroactively make that search protocol-compliant.

The search tool rewrote queries dynamically and only fetched three candidates per question, so this is a development pilot rather than the prespecified fixed-query search protocol. Unresolved means that this search and validator did not establish full support; it cannot establish that the public web lacks an answer. The protected Wikipedia index content hash and original paper-run item log are unavailable, so we cannot certify that a current page was independent of that protected snapshot or its downstream derivatives. The frozen rubric explicitly forbids a FULL verdict when that independence is unverified, for current as well as historical evidence. Consequently every protocol-level FULL outcome remains unresolved, including the 23 apparently answer-stating current pages. None of these judgments is an ACL decision or a measurement of a DP release.

Access-policy bookkeeping. The query class is unchanged when access changes. For each possible evidence route, a policy record identifies the principal (internal processor, external generator, or recipient), object (raw record or derivative release), operation (read, compute, publish, receive, or forward), scope, and ALLOW/DENY/UNKNOWN decision with provenance. The OPEN, RELEASE_ONLY, and SOURCE_AUTHORIZED routes summarized in Figure 1 are interface assumptions, not an ordinal ACL classification or access decisions inferred from the benchmark. An internal rule is K even if only a named employee may read it; a published case record is Case even if everyone may read it. A clinical population rate can be P and publicly available. Missing public evidence is distinct from an access denial or unknown policy, and an approved statistical release is separate from permission to expose raw source text to the generator.

G.2 What "required-evidence preservation" measures

For each query, an extractor first enumerated atomic facts and operations that the source reference made necessary for answering it. This list was produced from the reference SQL in ClinSQL and the reference answer in LegalQA; it is not the answer text itself. A separate call formed a query-conditioned typed object, and another call judged whether each required item could be read unambiguously from that object. Whole-request preservation is one only when every item passes. This measures whether entity, condition, outcome, scope, exception and operation remain linked, not final-answer accuracy or utility after DP noise.

The paired control contains the same terms, numbers, and authority strings as the typed object but removes field names and within-component links. It tests whether success came merely from retaining useful words. Missing verifier output and partial matches count as failure. We fixed prompts and parsing on 20 development items from each workload and did not change them during subsequent expansions. The final roster contains two primary 170-item ClinSQL strata and two 200-item LegalQA strata. The added ClinSQL stratum uses the test rather than validation split, and the added calls use singleton rather than five-item requests; we therefore report the strata separately and treat their union only as a post-hoc descriptive summary. All stages used GPT-5.4-mini. These labels are model-assisted, not expert gold or a human validation study.

A stored linked/unlinked pair. One saved ClinSQL query asks: "For my 68-year-old male patient, among men aged 63–73 what is the 75th percentile of distinct cardiac procedures per hospitalization?" Table 22 excerpts its stored representation and control, with the original model-assisted atom decisions. The source reference SQL, rather than independently audited clinical intent, defines the evidence requirements in this instrument.

A more complex ClinSQL example. One query asks for the 90th percentile of distinct procedures during the first 48 hours for female ICU patients aged 66–76, on their first ICU stay, with sepsis, and asks to compare hospital length of stay and mortality with age-matched controls. The required evidence includes the cohort filters, the 48-hour window relative to ICU admission, distinct procedure

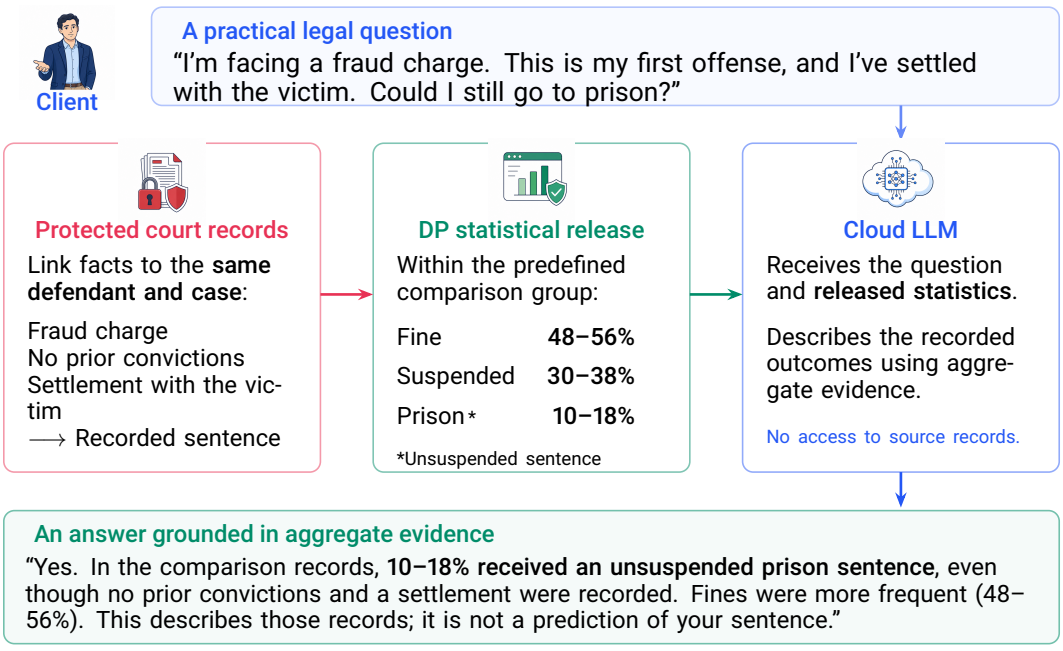


Figure 6: **Private statistics make a practical answer more informative.** Conceptual illustration with **invented numbers and an authored answer, not experimental results**. The question is generalized from a consultation entry in the local legal corpus; identifying details are omitted. A proposed, predeclared workload links defendant, settlement, and sentence before aggregation. The illustration assumes eligible single-defendant documents and a supplied public question-to-measurement binding; it does not illustrate an evaluated natural-language matcher. Public specification compilation occurs before protected measurement, and binding resolution is public post-processing. It assumes documents with resolved outcomes; each document is the counting and privacy unit. The intended DP release would support population descriptions without exposing individual records. These ranges have not been computed by the mechanism, and this fraud workload has not been evaluated. They neither estimate an individual’s risk nor measure the causal effect of settlement.

Table 22: Abridged fields from one actual paired representation. The complete unlinked control retains term counts and authority strings but removes field names and component links. Its percentile atom passes even though its join atom does not.

Required atom	Linked object excerpt	Unlinked control excerpt and decision
Join patients and procedures	subject: `procedures_icd`; predicate: “join on subject_id”; object: “patients”	`patients`: 5, `procedures`: 3, `subject_id`: 2; join atom missing
Count distinct codes	predicate: “count distinct”; condition: “group by subject_id”; object: `icd_code`	`distinct`: 4; counting-unit atom ambiguous
75th percentile	subject: “procedure_count distribution”; predicate: “75th percentile”; operation order: “compute patient counts then percentile across patients”	`75`: 2, `procedure_count`: 1; operator/-value atom preserved

identifiers per stay, the percentile operation, and joins among patient, admission, ICU-stay, diagnosis, and procedure records. A typed object keeps these roles and joins. An unlinked list containing “female,” “66–76,” “sepsis,” “48 hours,” “distinct,” and “90th percentile” does not specify which event anchors the window, what is counted, or the unit over which the percentile is computed.

LegalQA example. A joint-tortfeasor question requires several linked propositions: recourse arises after one obligor pays more than its share; allocation follows relative fault; that recourse claim is distinct from the victim’s original claim; and a particular Civil Act rule does not apply to the stated quasi-joint obligation. The order payment → recourse, the actor roles, and the negative

applicability condition are part of the evidence. Retaining the same legal terms without those links leaves several incompatible readings. This example is K under our taxonomy because the needed rules are person-independent, although the retrieved answer material may also contain a personal scenario.

G.3 Quantitative diagnostics

Table 23: Expanded model-assisted preservation audit. Zero-required rows and terminal schema failures are outside the assessable denominator. **Whole** ↑: requests for which every source-linked required item is preserved. **Atoms** ↑: individual required items preserved, shown as a partial-credit diagnostic only. *Relational* is the typed representation; *Unlinked* is the paired control retaining the same terms, numbers, and authority strings without field names or within-component links.

	Relational		Unlinked	
Workload	**Whole**	**Atoms**	**Whole**	**Atoms**
ClinSQL, prior stratum	139/158	1228/1257	35/158	658/1257
ClinSQL, added stratum	135/150	1529/1554	7/150	518/1554
ClinSQL, combined[†]	274/308	2757/2811	42/308	1176/2811
LegalQA, prior stratum	155/175	819/848	80/175	486/848
LegalQA, added stratum	142/152	911/925	21/152	318/925
LegalQA, combined[†]	297/327	1730/1773	101/327	804/1773

[†]Post-hoc union of the two primary strata. Selected-roster sizes were 340 ClinSQL and 400 LegalQA; 31/71 zero-required rows and 1/2 terminal schema failures, respectively, were not assessable.

Within the prior/added strata, the linked-minus-unlinked whole-request differences were 65.8/85.3 percentage points for ClinSQL and 42.9/79.6 points for LegalQA. For the post-hoc combined summaries, the paired differences were 75.3 ± 5.2 and 59.9 ± 5.5 percentage points (paired bootstrap 95% intervals), respectively. The differing split and request batching preclude interpreting the between-stratum change as improvement; the replicated claim is the within-stratum direction.

We also changed only the verifier family on a result-blind hash-min subset fixed during the earlier expansion. Two LegalQA rows had zero required atoms, leaving 58 assessable requests. The primary verifier passed 47/58 linked and 21/58 unlinked representations; an Anthropic-family verifier, shown the same GPT-produced atoms and candidate representations in hash-parity X/Y order, passed 30/58 and 2/58. Thus the paired direction remained positive (+44.8 and +48.3 percentage points), but condition-level agreement was only 63.8% for linked and 67.2% for unlinked, with agreement in both conditions 36.2%. Candidate format makes the nominally hidden conditions distinguishable, and neither verifier supplies independent gold. We therefore treat the direction as a sensitivity result and the absolute pass rates as verifier-dependent.

The legal retrieval audit used a fixed character 3–5-gram TF–IDF top-5 retriever as a probe, not as the paper's proposed retrieval system. It found a supporting payload for 67/80 requests. For 37/80, a supporting payload also contained a personal scenario; none contained a direct identifier, while 25 used pseudonymous or hypothetical roles and 10 contained contextual personal details. This establishes evidence mixing in this prototype corpus, not the rate of PII exposure in operational legal RAG.

G.4 Private RAG Interfaces Through the Taxonomy

The mappings below describe each interface's possible evidence path; they are not per-class utility claims made by the cited papers. An interface's ability to generate an answer format does not establish preservation of the evidence required for a correct answer.

Token- and logit-level aggregation. Grislain (2025) aggregates clipped, transformed document-conditioned log probabilities with a public-model prior and samples next tokens using an exponential mechanism. DPVoteRAG and DPSparseVoteRAG instead privatize token votes (Koga et al., 2025). These interfaces can convey K and express P/CP answers, but prediction agreement need not recover a corpus-wide count, denominator, or condition–outcome relation. Privacy cost depends on the specific mechanism and released steps: DPVoteRAG composes token releases, whereas DPSparseVoteRAG uses sparse-vector checks and a document-free model when predictions agree. K therefore does not imply an intrinsically large budget. The relevant comparison fixes total budget and request workload, or matches utility and reports the required budget and computation.

Table 24: Query–access taxonomy. Each route is available only when the source is sufficient and the operation and recipient are permitted. GEPI implements declared measurements in the DP-release column. A K support count measures document support, not a rule's truth. Aggregates cannot return a designated record.

	Evidence route		
Evidence need	**Open evidence**	**Approved DP release**	**Authorized source**
K Fact or rule	Public statement	DP support count for the stated fact or rule	Restricted fact or rule within scope
P Pattern across records	Published statistic	DP relational cohort measurement	Internal source aggregation
CP Pattern under stated conditions	Published matched comparison	DP conditional cohort measurement	Internal conditional aggregation
Case Contents of one record	The record, if public	—	The record within authorized scope

Keyword aggregation. DP-KSA aggregates tokens from document-conditioned answers into a private keyword output before final generation (Tang et al., 2026). This is evidence that a statistical intermediate can convey useful factual cues for K. Marginal keywords, however, do not specify the relations and denominators needed by P/CP. This RAG mechanism differs from the private in-context-learning KSA method of Wu et al. (2024), which would require adaptation here.

Synthetic corpora. DP-SynRAG releases synthetic text for later retrieval (Mori et al., 2026). It can retain general K content and support P/CP when generation preserves the relevant joint distribution and counting unit. A rate measured on synthetic documents is not automatically the source-corpus rate. Both a released synthetic corpus and a released synopsis can be reused through post-processing, so comparisons should include release construction and the same downstream request workload.

Relational statistics, Case, and compounds. A relational synopsis directly represents the counts and joint relations declared by a P/CP workload. It supports K only when required propositions fall inside a separately bounded family; document-support counts then mean documented support, not truth probability. Case facts remain an authorized record-retrieval problem because aggregate or synthetic evidence cannot certify what one requested record says. A compound request is supported only if every component is supported through its appropriate route.

G.5 Query-component annotation

Section 2 uses the frozen Query–Access v2.0.2 rubric, not the superseded preliminary taxonomy. The exact rubric is `workloads/query_access_k_audit_v2/RUBRIC_V2_0_2.md`; the 408-item extension is executed by `scripts/run_section2_query_components_408_taxonomy_v1.py`. The semantic pass receives the question, permitted user context, and public task contract, without a gold answer or retrieved protected passage.

Query–Access v2.0.2 annotation (abridged)

```
Task. Identify independently necessary answer components. K asks for facts, concepts, explicit
rules, or their application. P asks for empirical patterns across observational units. CP
applies empirical population evidence to a focal person's or case's conditions; an explicit
comparison group is unnecessary. Case asks for the contents or attribution of a specific
record. Separation. Public availability is not a semantic class and does not grant access.
User-supplied facts are not retrieved Case evidence. Treat CP as one component, not P plus CP.
Preserve unresolved ambiguity. Output. Record components, supporting query spans, rationale,
and any unspecified population, condition, outcome, operator, or scope fields. A P/CP label
does not establish that a count synopsis answers a causal request.
```

G.6 REQUIRED-EVIDENCE PRESERVATION

This instrument tests preservation, not answer correctness or privacy. It identifies required evidence before exposing either candidate representation; the paired comparison then checks every required atom.

Required-evidence preservation audit (three executed stages)

```
1. Identify required evidence. Identify only evidence atoms required to answer the query.
Use the reference solely as source-linked evidence; do not judge current legal or clinical
correctness. Include material denominator, unresolved-evidence, exception, ordering, and
authority requirements. Do not inspect any candidate representation.
2. Build the relational representation. Convert the supplied evidence into typed components
with subject, predicate, object, condition, exception, scope, authority, quantity, and
operation order. Preserve validity filters and multi-stage operations. This is a structural
representation, not a DP result or correctness judgment.
3. Judge preservation. Compare the required atoms with a typed relational object and an
object that retains only lexical terms, values, and authority strings. Decide whether
each representation alone preserves each atom without outside knowledge. A relation,
condition, exception, attribution, or order is missing when its words remain but their
linkage is ambiguous. Whole-query preservation requires every required atom and no material
contradiction. Do not fail the unlinked object automatically; pass facts that genuinely
require no linkage.
```

Each stage used a separate structured call, and the candidate representations were hidden while required evidence was identified.

H RECORD-COMPLETION EVALUATION

The target is whether a legal consultation documents prior criminal punishment. It is not independently verified criminal history. The main analysis uses 87 transfer targets, agent-adjudicated labels, 77 usable current-incident phrase cards and 10 explicit insufficient-card statuses. Cards include event class and broad region; phrases come from associated judgments in the same bundle. Shared bundle provenance does not by itself verify actor identity across sections. All methods use the same auxiliary clues, probability reader and threshold $p \geq .5$. The four complete conditions have 95 opposite-label pairs among 70 targets in 14 event–region cohorts. AUC contrasts are against auxiliary-only on common resolved targets, with 9,999 shared-target bootstrap draws. The main table reports the larger distance from the point estimate to either percentile bound as a conservative symmetric half-width. Brier loss, raw AUC, and failure states remain in the result artifacts. Labels and targets were exposed during the attack-evaluation lineage, so the analysis is secondary and outcome-informed, not a new untouched test.

Reader and release identity. The common hosted reader is OpenAI `gpt-5.4-mini-2026-03-17`. Its frozen instructions ask for the probability of documented prior punishment, use .5 when evidence gives no defensible direction, distinguish missing or conflicting evidence from NO, and forbid treating anonymous records as identity matches. The common supplement also warns that current-incident fragments may have uncertain roles or negation. Complete request bodies and the shared supplement are preserved in the record-completion manifest and runner listed in Appendix I.5. GEPI reuses the public shallow-275 release `legalm_shallow_275_gepi_eps4_20260923`: 415 coordinates, $B = 6$, $\sigma = 2.1793105272$, and calibration $(4, 1/3000)$. This is the same saved legal 4-label vector described in Appendix D.7, with standalone profile $(5.01, 10^{-5})$, not another draw. Attacking it adds no release cost; observing the separate 8-label utility release still requires the stated joint account.

Public-evidence adapters. GEPI and Tab-PE provide recipe-based evidence from released coordinates in the target's public event partition. Aug-PE supplies ranked released synthetic records. The historical packet interface caps each item at 700 characters and packs at most 6,000 excerpt characters; raw-context passes its complete released string, bounded by 128 sampled tokens. No adapter may inspect private records or evaluator labels. These channels do not exhaust all possible post-processing attacks.

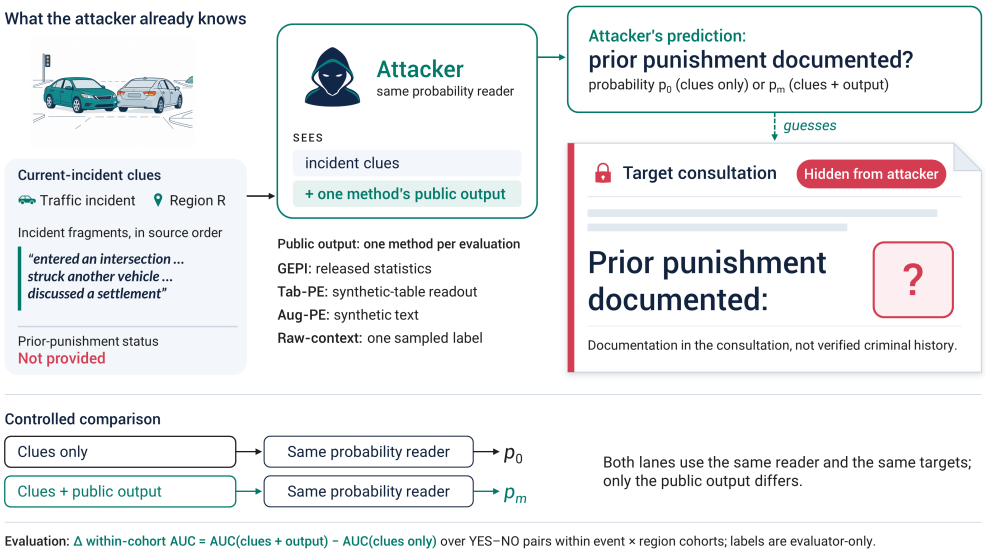


Figure 7: Record-completion attack and controlled comparison. The fictional traffic example illustrates the current-incident clues supplied to the attacker. For each target, the same probability reader sees either the clues alone or the clues plus one method's public output, then predicts whether the target consultation documents prior punishment. Evaluation compares within-cohort AUC over opposite-label pairs in the same event–region cohort; reference labels are evaluator-only. The target is documentation in the consultation, not independently verified criminal history. The illustrated phrases and prediction are explanatory, not measured attack results.

Final ordered-span attack. The final Table 2 attack uses every clue phrase in its frozen order and selects a record only when the ordered span is at most 1,500 characters. This rule is record-local and uses neither corpus rank nor evaluator labels. It retains 71/77 identified target records in an identity-only audit; 13/87 queries select no records and remain prior-only in the denominator. Each selected record supplies a transformed, centered three-label logit vector and the mechanism releases one sampled token. The transcript condition clips at ℓ_∞ radius .060893 and composes 87 events to $(3.99, 1/3000)$; the per-query condition clips at radius 2, giving pure $\epsilon = 4$ per query and $\epsilon \leq 348$ by basic composition. The matched $\epsilon = \infty$ condition removes clipping while keeping retrieval, crop, transform, public prior, and one-token output fixed. Their AUC changes relative to auxiliary-only are 0.0 ± 0.0, $+1.1 \pm 2.1$, and $+6.3 \pm 4.8$ percentage points, respectively (conservative 95% bootstrap half-widths; GEPI's is $+0.0 \pm 3.7$). The last interval excludes zero; the per-query interval includes zero. These observations qualify the adapter and demonstrate the cost of the transcript budget, not a general capability or privacy ordering for token-DP RAG. The $\epsilon = \infty$ condition remains matched to the public-fixture choice $\omega = .3$ and the same transform; the public prior still wins on 79/87 records, so it is a conservative endpoint for raw-context vulnerability. BA at the .5 threshold is uninformative here because the reader maps UNKNOWN to .5; the predeclared within-cohort AUC is primary.

Why the attack adapter changed. The earlier public-similarity utility adapter failed the population-evidence gate even without clipping. Its attack inputs contained the target gold span on 0/87 queries and selected the target record on 13/87; all 85 resolved predictions were .5. That null did not establish protection under usable record access. An earlier 16-token companion also changed the draw and cap, so it is not the matched $\epsilon = \infty$ condition reported here. The ordered-span rule above fixes the final interface before its new draws, while the full analysis remains outcome-informed as disclosed. Earlier missing outputs remain unresolved in their original records.

Explicit-evidence control. The hosted reader classifies 80/82 available non-private explicit-evidence controls correctly. This verifies sensitivity to clearly supplied documented attributes; availability selection and the evidence format distinguish it from a matched release condition.

Full-release stress test. A subsequent diagnostic indexes each entire saved $\epsilon = 4$ release. GEPI/Tab-PE expose all cells in the public partition, and Aug-PE uses BM25 top-5 over 600 syn-

thetic documents. One local Qwen2.5-7B reader produces 348/348 scores; all are NO at the fixed threshold, giving BA .500. Pair AUCs are .405/.595/.500/.574 for auxiliary-only/GEPI/Tab-PE/Aug-PE. The same scoring head reaches BA .980 on 82 gold-informed, availability-selected explicit-evidence controls under a different preamble. Those controls do not calibrate the attack prompt. GEPI/Tab-PE evidence is byte-identical within every opposite-label pair, so differences in their pair AUCs reflect clue–context interaction, not target-specific evidence. The reused targets, uncalibrated outcome-exposed scores, and one draw preclude a mechanism ranking.

Historical attack grid. The earlier 113-target multi-budget attack used coarse cards and a different reader; all nine intervals included zero. Its full rows are preserved in the historical artifact rather than pooled with the 87-target analysis.

H.1 Utility and attack channels

Table 25: Raw-context utility and final attack use different channels. Their observations are not a joint utility–attack trade-off on one answer path.

	Population utility	Ordered-span attack
Request	Supplied cohort/outcome population question	Documented prior-punishment attribute with auxiliary phrases
Record selection	Public-similarity adapter	Record-local ordered phrase span, at most 1,500 characters
Released object	Up to 128 sampled text tokens	One token from three transformed label logits
Downstream task	Generate a complete population answer	Read probability of the documented attribute
Budget scope	Stated utility/attack event reservation	87-event transcript, separate per-query condition, matched unclipped endpoint

Balanced accuracy measures decisions at $p = .5$; within-cohort AUC measures opposite-label ranking. The explicit-evidence control tests reader sensitivity, not attack optimality for each release.

I Reproducibility

The reproducibility bundle binds inputs, release objects, generation, judging and aggregation; superseded instruments are archived separately.

I.1 Scope

GEPI serves population requests supported by declared measurement bindings. Its privacy unit is one document or case bundle as defined in Appendix A.1. Source fidelity depends on document interpretation; the evaluation target is fidelity to frozen corpus states (Appendix A.4). Utility comparisons use fixed releases and block-specific judges. Method-specific privacy conditions are in Appendix E; the specified attack and its development history are in Appendix H.

Evaluation design. Table 1 uses common requests and block-specific references and scoring contracts. Original development and Qwen pilot parents are excluded. The medical roster retains 240 requests and adds 60. After observing abstentions on the retained pool, a common binding clarification was applied across the three evidence methods, without changing questions or numerical values. Scoring revisions are documented in Appendix F.7. The new medical requests and legal parents were evaluation-only at roster construction; subsequent reader revisions are recorded in Appendix I.5. All unresolved outcomes remain in completion bounds, including failures preceding output-cap amendments.

Availability. Medical source text is not redistributed without an authoritative upstream grant. Permitted artifacts, input identifiers and commitments support provenance checks. Exact evaluator references, source spans and private validation logs remain outside the DP release. Scored examples use public measurement requests and released or derived answer excerpts.

Table 26: Protocol conditions and the results they affect. Chronological logs and unused revisions are archived separately.

Condition	Affected interpretation	Location
Roster construction and protocol revisions	Common requests; retained-request repairs and subsequent reader revisions documented below	Appendices A.2, I.5
Medical v4/v5 and raw-score retention	Prompted caps differ; retained raw scores do not erase prompt effects	Appendix C.1
Medical release-label correction	Observed-after-scoring correction, not a preregistered condition label	Appendix D.7
Nominal baseline labels	Common column labels do not certify identical DP guarantees	Appendix E
Attack successor selection	Secondary, outcome-informed analysis with a specified reader and one draw	Appendix H

I.2 Conditions that change result interpretation

I.3 Detailed Release-and-Answer Algorithm

Algorithm 2 specifies input commitments, validation, privacy accounting and answer statuses. Stages A and D use public inputs and released outputs. Stage B and the exact sum in Stage C stay inside the data holder's trust boundary; only S is released. The optional precision policy propagates noise covariance and permits abstention. The main medical interval-task readout instead uses Equation 6 without that precision gate; the legal population task uses its executed known-eligible count readout. Individual evaluation adapters and prompts determine the evidence actually supplied.

I.4 Frozen Extraction and Source-Gold Validation

This source-gold validation protocol concerns the public medical relational control and ContractNLI, not the main LegalM/LeakRAG task-reference construction. The method separates measurement design, document interpretation, and corpus aggregation. Public benchmark material is used to define contracts and to test the document-local map g; it is not mixed with the protected corpus whose statistics will later be released. For each benchmark domain, a source adapter creates one immutable, label-free input per `doc_id`. A medical input is one declared encounter document; in the ContractNLI evaluation, a legal input is one supplied NDA source document. Rows or clauses from different identifiers are never joined, even when they share a patient, hospital, party, or source-case provenance field.

In the typed-fact path, the extractor does not directly supply the population count. It returns typed facts with exact source references and explicit completeness or uncertainty fields. A deterministic evaluator then applies the frozen contract to those facts to produce $(E_w(d), P_w(d))$. This factorization makes three failure locations observable: source evidence may be omitted or misread, the typed facts may be linked to the wrong contract condition, or the aggregation recipe may encode the document state incorrectly. Invalid schemas, missing required evidence, conflicting identifiers, and unresolved conditions are mapped to the declared unknown state rather than silently retried or converted to false.

The ContractNLI adapter instead predicts native document-NLI judgments and evidence directly; it is not an evaluation of the medical typed-fact interface. Their instructions are retained with the source-control artifacts; they are not the main population-generation prompts listed in Appendix I.5.

We validate this map with an independent source-gold adapter keyed by the same `doc_id`. The adapter derives reference typed facts and Boolean states from public database queries or native document annotations, without importing the predictor output. For relational medical contracts, the reference path stores the subject and encounter identifiers, both event rows, the strict time order, and the applicable window. For legal contracts, a reference judgment and an evidence span are retained separately: agreement with a benchmark label does not by itself establish that the cited clause supports the fixed meaning contract.

Algorithm 2 Detailed GEPI release-and-answer protocol

Require: Protected documents D; public workload $\mathcal{W}$ and selected profile; fixed local extractor g; family encoders $\{x_f\}$; fixed document-local partition rule κ; public recipes $\mathcal{R}$ and request bindings; noise scale σ; privacy ledger L; answer policy π

Ensure: Public release package S and answers derived without reopening D

A. Compile public specifications before the private mechanism

Validate contracts, parameter domains, unknown rules, and evidence requirements
Group only structurally identical contracts into families $\mathcal{F}$
Fix every coordinate, offset, encoder, and reconstruction recipe
Fix family partitions $\kappa(f)$; compute $m \leftarrow \sum_f m_f$ and $B \leftarrow \max_p \sum_{f:\kappa(f)=p} b_f$
Reject if a recipe leaves the fixed coordinates or an encoder can exceed B
Bind profile, extractor, schemas, recipes, and policy to versioned hashes

★ B. Interpret and encode each protected document locally

Check document identifiers; deduplicate identical copies and reject conflicts
$H \leftarrow \mathbf{0}_m$
for each remaining document $d \in D$ **do**
    $G_d \leftarrow g(d)$ using only d
    Validate schema and cited evidence; map unresolved required facts to U
    $p_d \leftarrow \kappa(d)$ using only d and the fixed rule
    for each selected family $f \in \mathcal{F}$ **do**
        if $\kappa(f) = p_d$ **then**
            Evaluate its frozen eligibility, relation, selector, and outcome rules on G_d
            Encode the resulting false, unknown, or eligible outcome state as $x_f(d)$
        else
            $x_f(d) \leftarrow \mathbf{0}_{m_f}$
        end if
    end for
    $x(d) \leftarrow \bigoplus_{f \in \mathcal{F}} x_f(d)$
    Reject before release unless $x(d) \in \{0,1\}^m$ and $\|x(d)\|_2^2 \le B$
    $H \leftarrow H + x(d)$
end for

C. Account for and publish one Gaussian synopsis

Ask L to reserve the charge for sensitivity $\sqrt{B}$ and scale σ; stop if denied
Draw fresh $Z \sim \mathcal{N}(0, \sigma^2 I_m)$ over all fixed coordinates
$\widetilde{H} \leftarrow H + Z$
Persist $S \leftarrow (\widetilde{H}, \mathcal{R}, \sigma,$ profile/policy metadata$)$
Commit the reserved charge; never publish H, $x(d)$, G_d, or private diagnostics

D. Answer each original request by post-processing S

for each original question q with user-supplied context c **do**
    Read the supplied public measurement bindings; do not define a new measurement
    for each supplied empirical measurement r **do**
        if r has no recipe in S **then**
            $\eta_r \leftarrow$ UNSUPPORTED
        else
            Apply r to $\widetilde{H}$; propagate Gaussian variance and covariance
            $\eta_r \leftarrow \pi(r(S)) \in \{$ANSWER, ABSTAIN$\}$
        end if
    end for
    Generate from unchanged (q, c), released evidence, $\{\eta_r\}$, and general knowledge
    Prevent general knowledge from filling a missing corpus statistic or record fact
    Validate evidence use, limitations, and output schema; otherwise return a public fallback
end for

Authorized record retrieval is a separate path and is outside this release guarantee.

Development and evaluation follow a fixed transaction:

1. choose document groups, contracts, metrics, and stopping rules using the declared development sources;
2. freeze the extractor code, prompt, model snapshot, label-free inputs, contract definitions, and their hashes;
3. execute once and seal raw responses, parsing failures, typed facts, and projected Boolean states before opening evaluation gold;

4. compute source-gold scores and the statistics in Equation 13; and
5. apply a predeclared audit-selection rule to source text, extracted evidence, and reference witnesses, then bind the result and audit receipts for network-free reproduction.

Development gates constrain both correctness and coverage. An extractor cannot pass merely by changing difficult decisions to unknown. If a gate fails, the corresponding evaluation split remains closed and a materially different extractor is treated as a new experiment. Outcome-stratified evidence audits diagnose selected positive, negative, and unknown cases; unless every case is reviewed, they are not estimates of full-population semantic accuracy.

Only after contracts, g, and the contribution map are fixed does protected execution evaluate each private document and form $H(D)$. Private records do not select workload items, vocabulary, thresholds, prompts, or retry policies in the stated mechanism. Any such adaptation is an additional data-dependent mechanism and requires a separate privacy analysis. Source-gold labels, document-level evidence, typed facts, and exact $H(D)$ remain internal; the public interface begins with the noisy synopsis in Equation 4.

Information-access boundary. Model placement depends on its inputs, not simply on whether it runs before or after release. A cloud model may help compile public source questions into candidate Boolean contracts using only public schemas and examples; deterministic checks admit only supported contracts, and the profile is fixed before protected access. Document interpretation and semantic evidence verification belong inside the data holder's trust boundary. Document-level judgments, extracted facts, evidence, and exact aggregates remain protected even when the original text is omitted. Schema checks, contribution bounds, aggregation, noise sampling, and accounting are implemented in code, with protected intermediates kept inside that boundary. After release, a cloud answer generator may receive only the request, the public contract and answer policy, and the noisy synopsis or its supported post-processed statistics. It cannot retrieve source documents or internal logs. This restriction preserves the post-processing guarantee; it does not guarantee that generated prose correctly describes the released statistics.

I.5 Executed Prompts and Reproducibility

This appendix records the instruments needed to interpret the reported results. Prompt excerpts are abridged; the listed frozen request bodies and code retain the complete instructions and output schemas.

Common-300 execution provenance. The common-roster campaign is bound by `scripts/run_table2_true300_unified_v1.py` and its frozen dispatch gate. Its reproducibility bundle is `workloads/table2_true300_builder_bundle_v1_20260926/`: it contains the roster, exposure audit, request–reference bindings, reuse matrix, release identities and all score rows. The saved aggregate contains the predeclared 10,000-draw cluster intervals and completion bounds. The independent audit checks all 84 cells and preserves output-cap recovery lineage. The constructors below identify the underlying instruments and their saved request bodies.

The Legal–Qwen GEPI $\epsilon = 1$ cell uses a typed C-only reader fixed after a 60-request diagnostic. We independently checked that reader on 60 additional parent–statistic pairs excluded from both the diagnostic and the final common-300 roster. Under the same saved release, generator route, seed, references and judge, the typed reader improved the mean score from 2.333 to 2.867 (paired difference $+0.533 \pm 0.213$, cohort-cluster 95% CI; 27/30/3 wins/ties/losses). The Table 1 run retains the original release and noise draw and scores 3.060 on 299/300 requests, with all-300 bounds $[3.053, 3.067]$. Full execution and transfer-audit artifacts are recorded in `workloads/legal_qwen_eps1_typed_readout_full300_v1_20260926/` and `workloads/legal_qwen_eps1_independent_transfer60_successor_v1_20260926/`.

The medical case-state extractor and its input reconstruction are defined in `scripts/medical_semantic_case_extractor_v6_0_1.py` and its pinned v6 dependency; the frozen 1,200-case state ledger is `workloads/medical_source_semantic_v6_0_1/case_states.json`. These extracted states define the main population references; they are not independent expert labels. Paths here are relative to the reproducibility repository. Medical population generation and v4 judging are constructed by `scripts/run_medical_external270_four_arm_v4.py` and its hash-pinned dependencies; the v4 rubric

is in scripts/run_medical_v264_gold_v4.py. The interval-cap correction and its exact changed request bodies are specified by scripts/plan_medical_external270_interval_cap_v5.py and retained under workloads/medical_external270_interval_cap_v5_20260924/. Legal population prompt constructors are in scripts/run_legal_external215_natural_frequency_v1.py; the legal Qwen generation and judge contracts are imported by scripts/run_legal_qwen38_fixed300_baseline_successor_v1.py and its budget-specific runners. Each run's frozen request bodies, rather than a generic prose-reference rubric, determine its exact input fields.

The main 87-target attack used OpenAI gpt-5.4-mini-2026-03-17. Complete instructions and output schemas are retained in workloads/legal_record_completion_attack_v1_20260925/requests.json; scripts/run_legal_record_completion_attack_v1.py records the common SUPPLEMENT. It warns that current-incident phrases may have uncertain roles or negation, that removed prior-status statements do not imply NO, and that synthetic or anonymous matches do not establish identity. The base instruction requires probability 0.5 when there is no directional evidence, permits applicable population counts to inform the known member's probability, and distinguishes unknown status from documented NO. The strict output fields are probability, answer, evidence_ids, and reason, with YES determined by probability ≥ 0.5.

The final raw-context ordered-span conditions are bound by scripts/run_raw_attack_ordered_span_private_v2.py and scripts/account_raw_attack_ordered_span_v1.py; their common hosted-reader execution is in scripts/run_raw_attack_ordered_span_reader_v1.py and workloads/raw_attack_ordered_span_reader_v1_20260926/RESULT.json. The saved plan binds the private-release inputs, common 87-target reader requests, and 9,999-draw within-cohort AUC analysis. These paths are distinct from the earlier public-similarity utility adapter.

Raw-context utility generation and judging use the constructors in scripts/run_raw_context_dprag_legal_table2_v5_private.py and scripts/run_raw_context_dprag_medical_table2_v3.py. scripts/run_raw_context_gpt_eps18_table2_v1.py reuses those contracts for the additional finite-budget cells. These are post-processing contracts for the sealed raw-context evidence.

I.6 Archived Instrument Development

Earlier consultation judges, LegalM-40 controls, population ladders and corpus revisions use superseded instruments and are retained in the reproducibility archive. They do not enter the common-300 results. The final scoring contracts and checks are specified in Appendix C.

I.7 Original-Consultation Control

A separate legal consultation-only diagnostic found no improvement from population evidence under its prose-reference endpoint: public-only and Aug-PE means were 4.547 and 4.543. A separate grounded-population instrument also did not separate GEPI and Tab-PE (-0.012 ± 0.037, 95% CI). These earlier readers and scoring contracts differ from the common-300 population tasks. They provide no evidence of a benefit for general consultation quality. Split construction, other methods, incomplete medical comparisons and superseded readouts remain in the reproducibility archive.

I.8 Post-Processing Fidelity

Release post-processing is checked separately by direct preservation of released estimates, intervals, uncertainty, and abstention. On the frozen LegalM $\epsilon = 8$ development simulation, generated prose preserved 28/31 released estimate ranges, 22/31 uncertainty ranges, and both for 20/31 ANSWER measurements across all routes. Separately, all 15/15 strict ANSWER question rows stated a group-not-individual limitation, and all 11/11 ABSTAIN measurements retained withholding. The 31 ANSWER measurements and 11 ABSTAIN measurements are distinct sets; the 15 strict ANSWER rows are question-level units, not a partition of 31. This is a post-processing-fidelity diagnostic, not answer correctness or a privacy claim. The prompt configuration that authored a reference is excluded from its ceiling condition to avoid scoring a model against its own output. We report no private-baseline score on this overlay.

I.9 Direct Endpoint Measurement

We evaluate the direct numeric interface of (6) on a separate instrument: 30 frozen development targets, one pre-committed noise draw per condition, and the evaluator's exact identified intervals as

Table 27: Held-out ContractNLI structured-release control (60 NDAs, 14 hypothesis clusters, 42 native-label rates). GPT is the uniformly repaired exposed-development score; MAE is against the exact native-label rate. Every row has 42/42 judgments. The non-private ceiling has no privacy claim.

Method	ϵ	GPT ↑	MAE ↓
GEPI, public prior	0	4.071	.0577
GEPI	4	4.071	.0525
GEPI, prior+DP successor†	4	4.167	.0470
GEPI, exact ceiling	∞	5.000	0
Tab-PE	4	3.000	.1738

the key. This is not the LegalM-40 population-evidence instrument, and the two sets of numbers are not comparable. Every condition returns an interval for all 30 targets, so nothing here depends on an abstention policy. **GoldAlign** ↑: one minus the mean endpoint loss against the evaluator's exact identified intervals. The released $\epsilon = 4$ condition attains a lower mean endpoint loss than the no-release condition, with GoldAlign at 0.99. The three-point GoldAlign ordering across $\epsilon = 0$, $\epsilon = 4$, and the zero-noise limit also holds, but it restates that same comparison together with the structural identity below and is not an independent result. Three qualifications apply. First, this is one draw on a development split and is descriptive only: we claim no trend in ϵ, and the $\epsilon = 1$ and $\epsilon = 8$ conditions are not placed on a ladder. Second, the zero-noise condition has zero loss by construction rather than by measurement, because the decoder applies to the noiseless aggregate the same rule that defines the exact key; it checks that the aggregation and decoding path is wired correctly and is not a point on a privacy–utility curve. Third, the reported score is computed against that key, which is derived from the protected corpus without differential privacy, so this scoring channel lies outside the accounting of Section 4.2, which covers the released histograms alone.

Medical source-consultation endpoint. We additionally run an authorized retrospective ceiling diagnostic on 263 paired, in-corpus questions from a leave-main-300-out split. Holding the generator and evaluator fixed, the non-private exact-interval condition ($\epsilon = \infty$) scores 4.958 on the 1–5 judge, versus 1.221 when the same packet contains no private aggregate ($\epsilon = 0$). The paired difference is 3.738 ± 0.092 (paired bootstrap 95% CI), with 261 wins, 2 ties, and no losses; all 526 judgments pass the frozen schema. The large difference primarily contrasts complete intervals with truthful abstentions. This is a same-source, in-corpus ceiling diagnostic, not an external held-out or finite-DP result, and therefore does not support a cross-domain main-method ranking or establish finite-ϵ medical utility.

I.10 Historical Population-Component Control

A historical 100-cell grid (20 cohort families by five outcomes) separated release availability from generated answers. The finite draw supported 28 cells and the zero-noise interface supported 45; the no-release condition abstained on all 100. Its earlier tolerance-based use score is not used here because that development instrument did not report its complete distance definition. The current numerical checks instead state their maximum-endpoint error and parser denominators explicitly (Section 5.3 and Appendix F.2).

I.11 ContractNLI Transfer Control

We therefore added a cleaner transfer control using ContractNLI test NDAs not present in either exposure boundary. After applying the exclusion rule fixed before opening the split, 60 NDAs remained; their 14 native hypotheses and three native labels define 42 population rates. Table 27 reports the one-shot result. All conditions answer all 42 rates. At matched $\epsilon = 4$, release MAE is lower than Tab-PE by 0.121 ± 0.052 (bootstrap CI). GEPI's MAE already decreases from .0577 to .0525 to 0 at $\epsilon = 0, 4, \infty$; the exposed prior+DP successor further reduces the $\epsilon = 4$ MAE to .0470. An audit found that the original runner passed each hypothesis token (e.g., `nda-10`) as the cohort even though the cohort is the common 60-NDA roster. We retain those scores only as invalid lineage and report a uniform exposed-development repair below: all five conditions receive the same deterministic cohort/hypothesis relabeling and corrected judge contract, while numbers, omissions, and uncertainty language remain frozen. This remains an oracle structured-release control—native annotations supply the protected table—not an end-to-end extraction result.

The † row is an exposed development successor, not a second private release or a held-out confirmation. It postprocesses the already released $\epsilon = 4$ vector by retaining the public train/dev prior and applying one hypothesis-level convex weight shared across its three labels. It requires no new protected-data access or privacy draw. On the same 42 rates it improves MAE relative to replacing the prior with the noisy vector, giving the desired MAE point ordering $\epsilon = 0 < 4 < \infty$. Because the rule was written after this 60-document result was exposed and uses only one DP draw, we report it provisionally and require an unexposed roster or seed before promotion. Under the repaired contract, successor minus public prior is $+0.095 \pm 0.286$ (14-hypothesis bootstrap CI); it does not separate. Successor minus Tab-PE is $+1.167 \pm 0.477$, while the original noisy-vector condition ties the public prior at 4.071. Thus the corrected diagnostic restores the desired point ordering for the successor but still does not isolate private-release value from the public prior.

I.12 Archived Mapping and Scale Diagnostics

The early 40-question LegalM development workload studied mapping coverage, semantic mismatch, cohort support and augmentation. It is distinct from the supplied measurement bindings in the common-300 evaluation. Its mapping errors and failed aggregate-error checks remain in the reproducibility archive; these diagnostics do not establish extraction accuracy for the final corpus. The final task construction and source-validation controls are in Appendices A.2 and I.4.

I.13 Generator Conditions and Requirement-Level Audit

The final generator comparisons appear in Table 1. The earlier transfer-parent table, its release identities, cap settings and exposure records remain in the reproducibility archive and are not pooled with the common-300 results.

A requirement-level audit of the earlier 300-request legal original-consultation diagnostic (Appendix I.7) explains why population evidence need not raise its consultation-reference score. Its references precede the appended population requirements used in Table 1. Across 6,846 mechanically retained gold sentences, the frozen conservative classifier finds 6,042 public-knowledge units, 2,409 query-stated case-fact units, and 58 individual-document units (labels are nonexclusive). It flags 5 possible population units; manual inspection of all five finds only case-specific negligence/share/-course ratios or generic practice tendencies, leaving 0 substantive corpus- population requirements. We report both counts rather than changing the rule after seeing its false positives. This is a source-interface audit rather than a correctness annotation, but it leaves no direct gold requirement for a population release to satisfy. A frozen 12-control calibration of the production 1–5 judge also passes all six predeclared method-neutral contrasts: equivalent numeric/verbal and plain/release-style answers tie at 5, while wrong-cohort and unjustified- certainty controls fall to 1. The observed mismatch therefore cannot be explained merely by the judge rewarding release-flavored wording.