QUERY-COUNT-INDEPENDENT DIFFERENTIALLY PRIVATE SYNOPSES FOR RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) can expose protected records through repeated access to its knowledge base. Query-time differential privacy (DP) requires cumulative accounting, which can constrain utility under a fixed total budget. Reusable DP synthetic corpora avoid per-query expenditure but face the challenge of preserving evidence for downstream retrieval. We propose SYNOPSISRAG, an offline-release framework that uses protected evidence to guide the private selection of intact public passages. A bounded fidelity score combines assessed answer completeness with estimated private-evidence support, guiding bidirectional reweighting and privacy-calibrated sampling of one representative per category. Online retrieval and generation then reuse the released synopsis without accessing the private corpus or incurring additional privacy loss for that version. We establish pure chunk-level DP with a direct document-level extension, derive conditional answer-support bounds for the released synopsis, and quantify how privacy calibration limits release concentration. On healthcare and finance workloads of up to 10,000 queries, SYNOPSISRAG achieves higher long-horizon match accuracy than the evaluated query-time and synthetic-corpus DP baselines. Extraction tests additionally show reductions in measured token leakage.

1 INTRODUCTION

Retrieval-Augmented Generation (RAG) enhances large language models (LLMs) by retrieving relevant information from an external knowledge base prior to generation. By grounding outputs in retrieved evidence, RAG can improve factuality, reduce hallucinations, and extend LLMs with domain-specific knowledge, making it a widely adopted paradigm for real-world LLM deployments.

In many practical settings, however, the knowledge bases used for retrieval consist of proprietary company-internal data, such as technical documentation, business records, or domain-specific expertise that directly underpins competitive advantage. Unlike public corpora, this data cannot be freely exposed during model interaction. Yet in standard RAG pipelines, the private datastore is repeatedly accessed at inference time, and the influence of individual records can surface in outputs in a query-dependent manner. Recent studies show that adaptive querying can exploit this interaction to infer sensitive content or membership in the datastore Jiang et al. (2024); Wang et al. (2026). As RAG systems are increasingly deployed as long-lived services, a central challenge is to protect the privacy of the retrieval database under continuous, adaptive, and potentially unbounded interaction, while preserving retrieval utility.

Differential privacy (DP) (Dwork, 2006) provides a principled framework for protecting sensitive data through formal guarantees quantified by a privacy budget ε. Recent work (Grislain, 2025; Koga et al., 2024; He et al., 2025; Wu et al., 2025) applies DP to RAG during query-time processing. Repeated private-data-dependent outputs require privacy accounting across interactions; under a fixed total budget, sustaining a long-lived service can therefore limit retrieval and generation utility. Alternatively, a system can release a differentially private corpus once and reuse it across subsequent queries. Such releases still face the challenge of preserving useful information under privacy constraints (Stadler et al., 2022; Giomi et al., 2022), particularly the evidence needed for downstream retrieval. These challenges motivate our central question: **how can we provide rigorous DP protection for long-term RAG interaction while preserving useful answer-bearing evidence?**

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.

To address this question, we propose **SYNOPSISRAG**, a *synopsis-based* DP framework that replaces online access to the private knowledge base with a compact collection of public passages. In the *offline* phase, a trusted curator uses evidence from protected records to assess how well public candidates support an anticipated query workload. A privacy-protected selection procedure produces category-conditioned synopses containing intact passages and their public provenance. Once released, all *online* retrieval and generation operate on these synopses. Subsequent adaptive interactions therefore incur *no additional privacy loss* for a fixed synopsis version.

The key idea is to use private evidence to guide the selection of public answer-bearing content. We first construct a query-independent passage pool from approved public resources. Public seed queries then guide the assessment of candidate passages, combining answer completeness with estimated support from private evidence. This construction preserves source wording while directing the limited synopsis capacity toward the anticipated workload. The released synopsis supports subsequent online interactions without further access to protected records. An optional offline refresh extension, presented in the appendix, accounts for additional releases under a lifetime privacy budget.

We summarize our contributions as follows:

- We formalize privacy-preserving RAG in the unbounded interaction regime and identify fundamental limitations of existing DP-based RAG methods under long-lived deployment.
- We propose SYNOPSISRAG, a retrieval-oriented DP synopsis framework that uses private evidence to select intact public passages for a one-time release, enabling unbounded online querying without additional privacy loss.
- We provide a theoretical analysis to provide provable privacy guarantees and characterize how synopsis approximation affects retrieval fidelity.
- We empirically demonstrate improved privacy–utility tradeoffs over query-time DP and synthetic database baselines, with up to $2\times$ higher match accuracy under 10,000 queries and up to 55.6% less token leakage under state-of-the-art inference attacks.

2 RELATED WORK

Privacy attacks on RAG. Prompt injection and adaptive extraction can induce a generator to expose retrieved content, while membership attacks infer whether a record is indexed. RAG-Thief (Jiang et al., 2024), Pirates of the RAG (Maio et al., 2024), instruction-based extraction (Qi et al., 2025), backdoor and worm-style attacks (Peng et al., 2024; Cohen et al., 2024), benign-query extraction (Wang et al., 2026; Hu et al., 2025), and membership inference (Zeng et al., 2024; Feng et al., 2025; Liu et al., 2025; Naseh et al., 2025; Yang et al., 2026; Chen et al., 2025) motivate protection of the observable interaction rather than only the database interface. DP addresses record influence on output distributions; it does not imply zero textual overlap or prevent disclosure of facts already public.

Private retrieval and generation. Koga et al. (2024) and Grislain (2025) study DP approaches to RAG. MURAG and MURAG-ADA use individual privacy filters and document-level accounting for multiple queries (Wu et al., 2025). Locally private entity perturbation protects a different object under a different trust model (He et al., 2025). Comparisons must therefore state adjacency, observable outputs, and the accountant, rather than equating mechanisms solely because they report the same numerical ε. Encryption, private information retrieval, and decentralized frameworks (Wang et al., 2025; E_Andersen et al., 2025) address access confidentiality, and random projection (Yao and Li, 2025) obfuscates embeddings without a formal guarantee; their guarantees are distinct from limiting information in answers.

Synthetic corpora and reusable releases. SAGE uses synthetic generation and refinement as an empirical privacy defense (Zeng et al., 2025); synthetic replacement alone is not DP (Stadler et al., 2022; Giomi et al., 2022), and related tools (Khan and Filkov, 2025; Mao et al., 2025) likewise lack a formal guarantee. DP-SynRAG explicitly constructs a reusable DP synthetic retrieval corpus (Mori et al., 2026). SYNOPSISRAG instead selects existing public text. This restriction makes the release domain data-independent, but limits the evidence that selection can convey. Reuse without additional privacy expenditure is the standard post-processing property, not a new privacy principle.

Private synopses and factual support. Continual-observation DP protects unbounded releases of fixed numerical statistics (Dwork et al., 2010a; Chan et al., 2011; Jain et al., 2023). Boosting-based private constructions emphasize poorly served queries while improving an approximation over a candidate space (Dwork et al., 2010b); DPI motivates our two-distribution fitting template in the setting of infinite data streams (Feng et al., 2024). Here the workload consists of natural-language questions and the score measures estimated answer-relevant support. MiniCheck (Tang et al., 2024) provides a local binary support estimator; its negative class does not separate contradictions from missing evidence, and its scores need not be calibrated probabilities. Our proof uses boundedness rather than checker accuracy.

3 PRELIMINARIES

3.1 PROBLEM SETTING

Domain-specific RAG. Let Ω specify the service's subject area, approved public sources, applicability metadata, and admissible query policy. Depending on the application, the metadata can include population, source date, reporting period, or jurisdiction. Let D_{pub} denote the approved public corpus and D_{priv} the multiset of indexed protected chunks. A fixed encoder $\phi(\cdot)$ maps text to embeddings, and a language model G produces an answer from a query and its retrieved context.

A conventional RAG system retrieves protected chunks for each query. Here, the curator constructs a synopsis S offline, and the serving system retrieves only from S. Public seed questions $\mathcal{Q} = \bigcup_{c=1}^{K} \mathcal{Q}_c$ describe the anticipated workload and guide candidate scoring and private fitting. The domain specification, seed workload, candidate text, model weights, and hyperparameters are fixed independently of D_{priv}. Medical and finance provide the experimental settings; law motivates the domain-scoped formulation but is not evaluated.

Privacy model. We assume a trusted curator that can read protected chunks and hold intermediate support scores. An adversary observes the released synopsis and subsequent answers, and may choose each query as a function of earlier outputs. Neighboring indexed corpora $D \sim D'$ differ by adding or removing one chunk. A mechanism $\mathcal{M}$ is pure ε-DP if, for every neighboring pair and measurable event O,

$$\Pr[\mathcal{M}(D) \in O] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in O]. \tag{1}$$

We apply this definition to the synopsis and the adaptive answer transcript. User queries are inputs to the service and are not the protected records in this definition. Appendix F establishes document-level protection using the same bounded-update calibration.

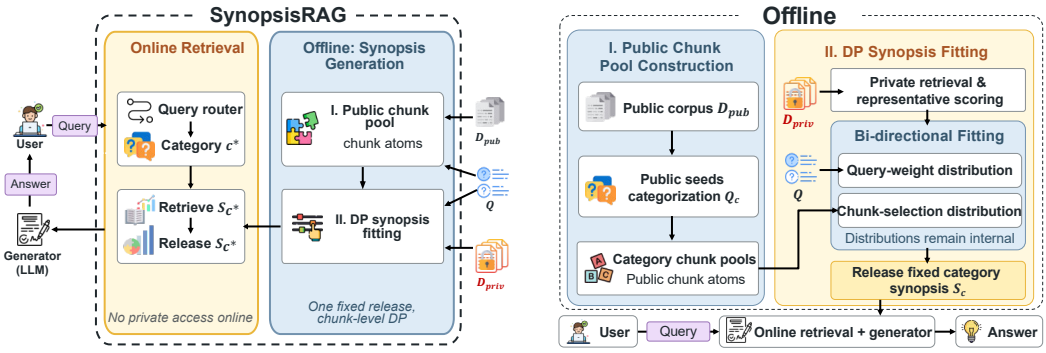
Utility objective. We aim to select a compact set of public passages that supports the domain's anticipated questions and reflects evidence in the protected corpus. Online answering uses that set without accessing protected chunks. We assess the private guidance against a public-only selection with the same candidate space and release capacity, thereby separating the contribution of fitting from public pool construction.

3.2 PUBLIC SCOPE AND SYNOPSIS REPRESENTATION

A synopsis is a reusable data representation from which query answers can be approximated (Dwork et al., 2010b). DPI constructs data-independent candidate synopses and privately adapts their sampling weights for streaming queries (Feng et al., 2024). In our RAG setting, the synopsis consists of selected public passages, identifiers, and source metadata. Each passage belongs to a public category, which determines its offline release quota. For each admissible query, the public router selects one category, and the generator receives that category's complete released passage and public source metadata. The system maintains a distinction between domain membership and available evidence: a public query policy handles out-of-scope requests, while the retrieval and generation stages handle accepted queries using the released context. Neither decision initiates private retrieval. The development evaluation uses domain-specific question sets and does not assess open-set routing.

3.3 BIDIRECTIONAL BOOSTING AS A FITTING STRATEGY

Multiplicative-weight updates underpin classical boosting (Freund and Schapire, 1997). Boosting-based private synopsis constructions emphasize poorly served queries (Dwork et al., 2010b); DPI



(a) Overview of SynopsisRAG.

(b) Offline synopsis generation.

Figure 1: SynopsisRAG pipeline for offline DP synopsis generation and online retrieval. (b) details the offline synopsis-generation component in (a).

motivates bidirectional updates over query and candidate-synopsis distributions (Feng et al., 2024). We use this fitting template with bounded answer-support feedback. Candidate updates favor higher-scoring passages, and query updates emphasize poorly served seeds. Section 4.1 specifies the simultaneous updates and final randomized release; Section 4.3 establishes privacy for that sampled output.

4 SynopsisRAG

Overview. SynopsisRAG decouples privacy expenditure from online RAG interaction by confining private-data access to an offline synopsis-generation phase. The curator spends a global privacy budget to release one public representative per category, and the resulting synopsis is reused for subsequent queries. Online routing and generation depend only on the query, public configuration, and released passages. By the post-processing property of DP (Dwork and Roth, 2014), adaptive online interactions therefore incur no additional privacy loss for a fixed release. This design places privacy accounting at synopsis construction and allows the serving system to operate without access to the protected corpus.

The utility challenge is to select public passages that both answer the anticipated questions and agree with applicable private evidence. We use intact public chunks as candidates to preserve readable source text and provenance while fixing the possible release contents independently of the private corpus. Query-independent chunking and public categorization produce the candidate pools and routing metadata. Within these pools, answer-bearing retrieval connects each query–candidate pair to relevant private evidence. A bounded fidelity score combines the completeness of the candidate's answer-bearing content with estimated support from that evidence. This gives private fitting a signal tied to the usefulness of the passage as answer context.

To guide selection across the seed workload, the curator maintains two internal distributions: a candidate distribution that favors passages with higher feedback and a query distribution that emphasizes poorly served questions. Simultaneous exponential reweighting lets these distributions shape subsequent fitting rounds. The bounded feedback permits privacy calibration of the final representative sample, including the influence of both data-adaptive distributions (Theorem 1). Only the sampled public passage and its public provenance are released; scores and fitting weights remain inside the curator.

Figure 1 summarizes the workflow. Phase I constructs public candidate pools; Phase II scores and fits the candidates using protected evidence, then samples one representative per category. Online, the public router selects a category and supplies its complete released passage to the generator. Algorithm 1 specifies the offline construction. Optional refreshes and their additional privacy accounting are presented in the appendix.

4.1 OFFLINE: SYNOPSIS GENERATION

Phase I: privacy-free public chunk pool construction. We split D_{pub} into source-linked chunks using fixed sentence boundaries and chunking settings. Each atom a is an intact public chunk identified by its source and offsets. The resulting collection P_{pub} contains the chunks of all public documents with immutable source identifiers.

The eligible chunks are selected from P_{pub} using fixed public source and metadata policies. We group these chunks by public chunk embeddings or declared domain labels, assigning each chunk to one category to form the candidate pool $\mathcal{A}_c$. Category centroids in the same encoder space define the public query router. The finite pools and router are constructed from public documents and fixed public configuration before seed-query scoring and private access. For the population guarantee, they are also fixed independently of the seed sample used to estimate risk.

Phase II: retrieval-aligned DP synopsis fitting. For each seed query $q \in Q_c$ and public candidate $a \in \mathcal{A}_c$, FLAN-T5-XL (Chung et al., 2024) selects the answer-bearing source sentences $C(q,a)$ and assesses their answer completeness $b(q,a) \in [0,1]$. We use standard retrieval, conditioned on the query and selected answer text, to obtain k private evidence windows, denoted $N_k(D;q,a)$. MiniCheck (Tang et al., 2024) then scores each selected sentence against each retrieved window. Appendix B.1–B.2 specifies the checkpoints, extraction prompt, retrieval rule, and scoring interface.

Our bounded fidelity score combines answer completeness with the strongest applicable window's support for all selected sentences:

$$f_{q,a}(D) = \begin{cases} b(q,a) \max\limits_{d \in N_k(D;q,a)} \left[\chi(q,a,d) \min\limits_{s \in C(q,a)} M(q,d,s) \right], & \substack{C(q,a)\neq\varnothing,\\ N_k(D;q,a)\neq\varnothing,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Here $M(q,d,s) \in [0,1]$ is the MiniCheck support score, and $\chi(q,a,d)$ checks the declared entity, population, period, and scope compatibility. The minimum requires joint support within one window; the maximum selects the strongest window. Empty selections or empty retrieval results receive score zero.

We map fidelity to bounded feedback with target τ and margin μ:

$$s_{q,a}(D) = \left[\frac{2(f_{q,a}(D) - (\tau - \mu))}{\mu} - 1 \right]_{-1}^{1}, \qquad 0 < \mu \le \tau \le 1, \tag{3}$$

where $[x]_{-1}^{1} = \min\{1, \max\{-1, x\}\}$. Feedback increases linearly from -1 at $\tau - \mu$ to $+1$ at τ, with clipping outside this interval. The curator computes the feedback table once per fixed snapshot for the reweighting stage.

We maintain a synopsis distribution $P_{s,c}$ over candidate chunks and a query distribution $P_{q,c}$ over seed queries. For each active category, we initialize $P_{s,c}^{(1)} = \text{Unif}(\mathcal{A}_c)$ and $P_{q,c}^{(1)} = \text{Unif}(Q_c)$. For a fixed public number of rounds $t \ge 1$, each round $i = 1, \dots, t$ computes the candidate's average feedback and the query's average negative feedback from the current distributions:

$$g_i(a) = \sum_{q \in Q_c} P_{q,c}^{(i)}(q) s_{q,a}, \qquad h_i(q) = -\sum_{a \in \mathcal{A}_c} P_{s,c}^{(i)}(a) s_{q,a}. \tag{4}$$

Both functions lie in $[-1,1]$. Update both distributions simultaneously:

$$P_{s,c}^{(i+1)}(a) = \frac{P_{s,c}^{(i)}(a) e^{\alpha_{s,i} g_i(a)}}{\sum_{b \in \mathcal{A}_c} P_{s,c}^{(i)}(b) e^{\alpha_{s,i} g_i(b)}}, \tag{5}$$

$$P_{q,c}^{(i+1)}(q) = \frac{P_{q,c}^{(i)}(q) e^{\alpha_{q,i} h_i(q)}}{\sum_{q' \in Q_c} P_{q,c}^{(i)}(q') e^{\alpha_{q,i} h_i(q')}}. \tag{6}$$

The synopsis update increases the weight of chunks with higher support feedback. The query update increases the weight of poorly served seeds, focusing subsequent rounds on those queries. Both distributions and the feedback table are maintained inside the curator throughout the fixed fitting schedule.

Since $s_{q,a} \in [-1, 1]$, every candidate update satisfies $g_i(a) \in [-1, 1]$ under any private corpus. We calibrate the final sampled representative by setting

$$\alpha_{s,i} = \alpha_{q,i} = \frac{\varepsilon_c}{4t}, \qquad i = 1, \ldots, t. \tag{7}$$

Theorem 1 establishes the privacy guarantee for this schedule, including the private-data dependence of both fitting distributions. After fitting, release exactly one representative

$$S_c \sim P_{s,c}^{(t+1)}, \qquad S_c \in \mathcal{A}_c. \tag{8}$$

The release consists of the sampled complete source passage, its public identifier, offsets, and metadata, together with the public router. The final categorical sample provides the privacy-protecting randomness. A category with $\varepsilon_c = 0$ uses a sample from the public initial distribution. An empty candidate pool gives a fixed insufficient-evidence fallback.

4.2 Online Retrieval

For an admissible query q, the public router chooses category $c^*(q)$, and we retrieve its released representative $S_{c^*(q)}$. The generator receives the query and the complete original passage with its public source metadata. The passage fits the evidence budget B established during public construction; separate prompt space is reserved for instructions, the query, and the response. A public query-length rule rejects oversized requests.

The generator is instructed to ground its response in the supplied evidence and acknowledge insufficiency. An absent or currently disallowed passage yields a fixed insufficient-evidence response. The online service operates on queries, public configuration, and released chunks, so answering queries is post-processing of the synopsis. The primary method reuses a fixed synopsis throughout deployment. Appendices B.4 and E describe an optional offline refresh extension and its privacy accounting.

4.3 Theoretical Analysis

We analyze exact arithmetic and exact categorical sampling conditional on each release's public configuration. All step sizes are nonnegative and fixed before that release's private access.

Theorem 1 (Privacy of a sampled category synopsis)**.** *Fix the public candidate and query supports, public initialization, and nonnegative step-size sequences. With feedback in $[-1, 1]$, one representative sampled from the final candidate distribution is pure ε_c-DP whenever*

$$4 \sum_{i=1}^{t} \alpha_{s,i} \le \varepsilon_c. \tag{9}$$

In particular, Eq. (7) satisfies this condition. Category releases compose to pure $\sum_c \varepsilon_c$-DP, and their adaptive online answer transcript has the same bound at every finite horizon.

The proof uses $g_i(a) \in [-1, 1]$ to bound changes in the cumulative exponent and its normalization constant. It analyzes the released representative directly, including the influence of the private query weights. Appendix C provides the proof.

Theorem 2 (Average answer-support bound)**.** *Fix a category, corpus, and seed workload, and let $P = P_{s,c}^{(t+1)}$. Define the private analysis quantities*

$$\widehat{\ell}_{\mu,c}(P) = \mathbb{E}_{q \sim \mathrm{Unif}(Q_c),\, a \sim P} \frac{1 - s_{q,a}}{2} = \frac{1 - \Gamma_c(P)}{2}, \quad \Gamma_c(P) = \mathbb{E}_{q,a}[s_{q,a}]. \tag{10}$$

For a single draw $S_c \sim P$,

$$\Pr_{q \sim \mathrm{Unif}(Q_c),\, S_c \sim P}[f_{q,S_c} \le \tau - \mu] \le \widehat{\ell}_{\mu,c}(P). \tag{11}$$

If, additionally, Q_c consists of m_c i.i.d. draws from a deployment distribution Π_c, independently of the private corpus and the fixed public pool, router, and scoring contract, then with probability at least $1 - \rho$ over Q_c,

$$\Pr_{q \sim \Pi_c,\, S_c \sim P}[f_{q,S_c} \le \tau - \mu] \le \min\left\{1, \widehat{\ell}_{\mu,c}(P) + \sqrt{\frac{\log(2|\mathcal{A}_c|/\rho)}{2m_c}}\right\}. \tag{12}$$

The bound characterizes average low-support risk over the workload and the one-time release draw for any fitted distribution. Equation (10) defines internal analysis quantities. The population extension requires the stated i.i.d. sampling assumption; curated or adaptively buffered seeds retain the empirical interpretation. Appendix D proves the result uniformly over distributions on the fixed public pool and discusses its scope.

Corollary 1 (A realized synopsis)**.** *For $0 < \eta < 1$, let $R_c(a) = \Pr_{q\sim\mathrm{Unif}(Q_c)}[f_{q,a} \le \tau - \mu]$. With probability at least $1 - \eta$ over the single release draw,*

$$R_c(S_c) \le \min\{1, \widehat{\ell}_{\mu,c}(P)/\eta\}. \tag{13}$$

Markov's inequality gives this weaker, possibly vacuous realization guarantee. The analogous population statement uses the right side of Eq. (12) in place of the empirical risk and has joint failure probability at most $\rho + \eta$. The following proposition states the additional semantic error needed to relate these score bounds to valid answer-bearing evidence.

Proposition 1 (Conditional answer-bearing evidence bound)**.** *Let $W(q, a) \in \{0, 1\}$ be an independently specified label indicating that passage a provides a sufficient, correct, and applicable answer to q. For the fixed category, corpus, workload, and final distribution P, define*

$$\xi_c(P) = \Pr_{q\sim\mathrm{Unif}(Q_c),\, a\sim P}[f_{q,a} > \tau - \mu,\ W(q, a) = 0]. \tag{14}$$

Then the probability that the sampled representative supplies no valid answer-bearing evidence obeys

$$\Pr_{q,S_c}[W(q, S_c) = 0] \le \min\{1, \widehat{\ell}_{\mu,c}(P) + \xi_c(P)\}. \tag{15}$$

The term $\xi_c(P)$ captures semantic errors among high-scoring chunks, including missed qualifiers, scope errors, and false support judgments. Its magnitude requires independent validity assessment. Both terms are internal analysis quantities; publishing their evaluations requires separate justification. The bound concerns the complete passage served under the analyzed public configuration. Correct generation additionally depends on the generator. Appendix D proves the event decomposition and a corresponding realized-synopsis statement.

5 Experiments

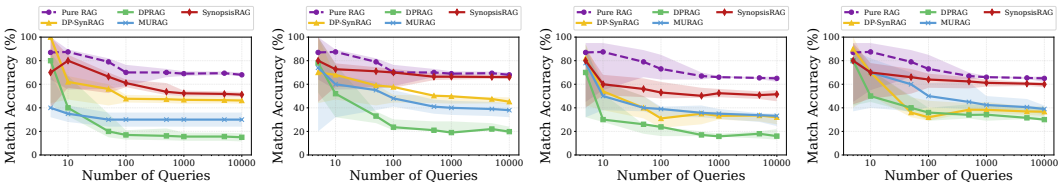
We evaluate SynopsisRAG on healthcare and finance tasks using two language-model backbones. The experiments measure answer utility over long query sequences, robustness to workload and public–private vocabulary mismatch, exposure under extraction attacks, and online retrieval latency. Each primary SynopsisRAG run uses the answer-bearing fidelity score and bidirectional fitting procedure of Section 4.1, releases one complete public representative per category, and serves queries using the category-routing procedure of Section 4.2. The query-volume experiments reuse a fixed synopsis throughout the evaluated sequence.

5.1 Setup

Hardware and models. All experiments run on a machine equipped with an NVIDIA A100 GPU and 64 GB of RAM. We use LLaMA-2-7B and Mistral-7B as answer generators in separate evaluations. The offline scoring models and prompts are specified in Appendix B.

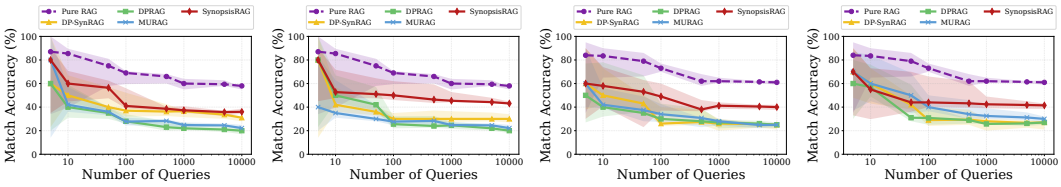
Datasets. Each domain uses a protected corpus, a public candidate corpus, evaluation questions, and public category resources. For healthcare, HealthCareMagic-100k (Zichuan) serves as the protected corpus, Medical Wikipedia (Varma et al., 2021) supplies public candidate passages, and Comprehensive Medical QA (Sabado) supplies evaluation questions. MedQuAD (Ben Abacha and Demner-Fushman, 2019) provides the public category resources. For finance, the corresponding resources are IndustryCorpus_finance (BAAI), Wikipedia Finance Articles (Foly et al., 2025), FIQA (Rahib), and FinanceQA (Choubey), respectively.

Default parameters. Unless otherwise stated, we use $K = 10$ public categories, $t = 2$ fitting rounds, and $k = 3$ private evidence windows per seed–candidate pair, with feedback target $\tau = 0.7$ and margin $\mu = 0.4$. The main comparisons use total synopsis-release budgets $\varepsilon \in \{1, 10\}$, allocated equally across categories: $\varepsilon_c = \varepsilon/K$. Both fitting distributions use step sizes $\alpha_{s,i} = \alpha_{q,i} = \varepsilon/(4Kt)$.



(a) LLaMA-2-7B, $\varepsilon = 1$ (b) LLaMA-2-7B, $\varepsilon = 10$ (c) Mistral-7B, $\varepsilon = 1$ (d) Mistral-7B, $\varepsilon = 10$

Figure 2: Match accuracy versus query count on healthcare across models and privacy budgets.



(a) LLaMA-2-7B, $\varepsilon = 1$ (b) LLaMA-2-7B, $\varepsilon = 10$ (c) Mistral-7B, $\varepsilon = 1$ (d) Mistral-7B, $\varepsilon = 10$

Figure 3: Match accuracy versus query count on finance across models and privacy budgets.

Each synopsis contains one sampled public representative per category and remains fixed throughout the query-volume evaluation, without offline refreshes.

Baselines. We compare against DPRAG (Grislain, 2025), MURAG (Wu et al., 2025), DP-SynRAG (Mori et al., 2026), and Pure RAG. DPRAG provides pure DP at query time, while MURAG supports multiple queries through individual per-document privacy filters. DP-SynRAG instead constructs a reusable differentially private synthetic retrieval corpus. Pure RAG retrieves directly from the protected corpus without a privacy mechanism, serving as a nonprivate utility reference. **Attacks.** We evaluate extraction exposure using RAG-Thief (Jiang et al., 2024), Pirate (Maio et al., 2024), GEA (Cohen et al., 2024), and PIDE (Qi et al., 2025). These attacks attempt to induce the system to disclose datastore content through adversarial queries or injected instructions. Appendix H describes the attacks.

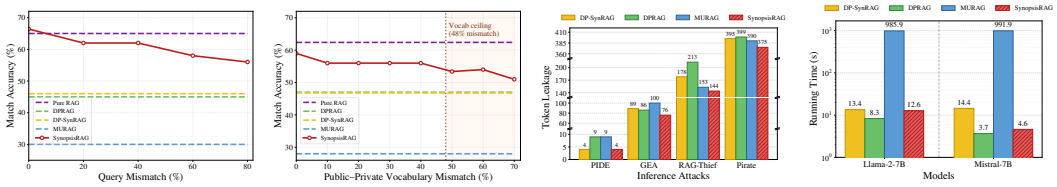
Metrics. We report three metrics. *Answer utility* is measured by match accuracy, comparing generated answers with reference answers using the matching procedure of Wu et al. (2025). *Empirical extraction exposure* is measured by total stolen tokens, computed from token-level overlap between attack-extracted text and the protected corpus (Jiang et al., 2024; Maio et al., 2024). This attack-based metric is distinct from the formal DP guarantee. *Online efficiency* is measured by average per-query retrieval latency, excluding offline synopsis construction: public pool construction, support checking, and fitting.

5.2 Results

Utility. Figures 2 and 3 report match accuracy on healthcare and finance as the evaluated query count increases. SynopsisRAG maintains a long-horizon accuracy advantage over the evaluated DP baselines. On healthcare with LLaMA-2-7B at $\varepsilon = 1$, accuracy at 10,000 queries is 51% for SynopsisRAG, 68% for Pure RAG, 46% for DP-SynRAG, 18% for DPRAG, and 30% for MURAG. This corresponds to gains of 5%, 33%, and 21% over the three DP baselines, respectively, while remaining 17% below Pure RAG. The finance results show a similar long-horizon advantage, indicating that the observed benefit extends across both evaluated domains.

The synopsis remains fixed throughout each run, so changes in its measured accuracy across query counts do not arise from additional privacy expenditure. Small query sets can produce variable accuracy estimates because individual questions have a larger influence on the reported average. Evaluating more queries changes the evaluated sample and can yield more stable estimates for a fixed synopsis. The sustained performance therefore demonstrates the utility of reusing the released public context over long query sequences, while variability across independently sampled synopses remains a separate consideration.

Robustness to query mismatch. Figure 4a evaluates robustness to differences between the seed workload and evaluation queries. We construct mismatch by randomly replacing 0%–80% of seed-



(a) Accuracy under query mismatch.

(b) Accuracy under domain mismatch.

(c) Token leakage under different attacks.

(d) Average retrieval time per query.

Figure 4: Robustness, privacy, and efficiency on healthcare.

aligned evaluation queries with queries from topics not covered by the seed workload, keeping the evaluation-set size and released synopsis fixed. Match accuracy decreases from 66.4% to 56.0%, a decline of 10.4 percentage points. This result indicates that workload alignment affects synopsis utility, although useful answering performance is retained under substantial mismatch.

Robustness to vocabulary mismatch. Figure 4b evaluates robustness to public–private vocabulary mismatch, with the dashed line marking the observed 48% mismatch in healthcare. We construct mismatch by randomly removing vocabulary shared with the private corpus from the public corpus to obtain target mismatch levels between 0% and 70%. Match accuracy generally decreases as mismatch increases, with lower accuracy at the higher mismatch levels. This result indicates that lexical alignment contributes to synopsis utility, while answer-bearing coverage also depends on the facts and qualifications retained in the public atoms.

Privacy. Figure 4c reports tokens leaked under each attack. SYNOPSISRAG achieves the lowest leakage across all four. Under RAG-Thief and Pirate it leaks 144 and 375 tokens, the largest reductions relative to DPRAG (69 and 24 fewer tokens); under PIDE it leaks 4 tokens against 9 for DPRAG and MURAG (55.6% lower). Empirical attacks complement but do not establish DP; the generator never sees private chunks, so less extractable content is exposed at the same nominal budget.

Efficiency. Figure 4d reports average per-query retrieval time. With Mistral-7B, SYNOPSISRAG takes 4.59 s per query, versus 3.66 s for DPRAG, 14.44 s for DP-SynRAG, and 991.9 s for MURAG. With LLaMA-2-7B, it takes 12.6 s, versus 8.3 s for DPRAG and 13.4 s for DP-SynRAG. These results show that SYNOPSISRAG reduces online cost relative to several baselines but is not uniformly the fastest. Offline construction and fitting are excluded.

Ablations. We examine (i) fitted versus uniform release, (ii) no-retrieval and public/private RAG references, and (iii) sensitivity to ε, K, and k. Figure 5 presents the healthcare comparisons. At 500 queries with LLaMA-2-7B, fitted releases achieve 54% and 67% match accuracy at $\varepsilon = 1$ and 10, respectively, versus 50% for uniform release. Table 2 further evaluates the effects of the privacy budget ε, category count K, and private retrieval depth k. Additional finance comparisons and detailed ablation results are provided in Appendix I.

6 CONCLUSION

In this work, we propose SYNOPSISRAG, a synopsis-based differentially private framework that decouples privacy expenditure from the number of online interactions. Using bounded answer-completeness and evidence-support scores, the framework privately selects intact public passages through exponential reweighting and a sampled release. Online retrieval and generation reuse the fixed synopsis without additional privacy expenditure. The appendix provides privacy accounting for optional offline refreshes. Our theoretical analysis establishes privacy guarantees and characterizes answer-support bounds and limits on release concentration, while experiments demonstrate improved privacy–accuracy tradeoffs over the evaluated query-time DP and DP synthetic-database baselines. The framework's utility depends on public-pool coverage, the reliability of answer selection and support assessment, and alignment with the query workload. In particular, selection cannot supply proprietary facts absent from the public candidates, and estimated evidence support does not guarantee answer correctness. The method also assumes access to suitable domain-specific public resources; extending it to mixed-domain workloads and improving evidence coverage under a fixed privacy budget are directions for future work.

REFERENCES

BAAI. IndustryCorpus_finance. https://huggingface.co/datasets/BAAI/IndustryCorpus_finance.

Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23, 2019.

T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Trans. Inf. Syst. Secur.*, 14(3):26:1–26:24, 2011.

Yufei Chen, Yao Wang, Haibin Zhang, and Tao Gu. Fine-grained privacy extraction from retrieval-augmented generation systems via knowledge asymmetry exploitation. *CoRR*, abs/2507.23229, 2025. doi: 10.48550/ARXIV.2507.23229.

Amit Choubey. FinanceQA. https://huggingface.co/datasets/sweatSmile/FinanceQA.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of machine learning research*, 25(70):1–53, 2024.

Stav Cohen, Ron Bitton, and Ben Nassi. Unleashing worms and extracting data: Escalating the outcome of attacks against RAG-based inference in scale and severity using jailbreaking. *CoRR*, abs/2409.08045, 2024. doi: 10.48550/ARXIV.2409.08045.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

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

Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential privacy under continual observation. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 715–724. ACM, 2010a.

Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, Las Vegas, Nevada, USA, October 23-26, 2010*, pages 51–60. IEEE Computer Society, 2010b.

Tessa E_Andersen, Ayanna Marie Avalos, Gaby G_Dagher, and Min Long. D-rag: A privacy-preserving framework for decentralized rag using blockchain. 2025.

Kaiyue Feng, Guangsheng Zhang, Huan Tian, Heng Xu, Yanjun Zhang, Tianqing Zhu, Ming Ding, and Bo Liu. RAGLeak: Membership inference attacks on rag-based large language models. In Willy Susilo and Josef Pieprzyk, editors, *Information Security and Privacy - 30th Australasian Conference, ACISP 2025, Wollongong, NSW, Australia, July 14-16, 2025, Proceedings, Part III*, volume 15660 of *Lecture Notes in Computer Science*, pages 147–166. Springer, 2025.

Shuya Feng, Meisam Mohammady, Han Wang, Xiaochen Li, Zhan Qin, and Yuan Hong. DPI: ensuring strict differential privacy for infinite data streaming. In *IEEE Symposium on Security and Privacy, SP 2024, San Francisco, CA, USA, May 19-23, 2024*, pages 1009–1027. IEEE, 2024.

Sabine Foly, Jingshu Liu, Jean-Gabriel Barthelemy, Gaëtan Caillaut, Raheel Qader, Mariam Nakhle, and Arezki Sadoune. Wikipedia-Finance-Articles, 2025. URL https://huggingface.co/datasets/DragonLLM/Wikipedia-Finance-Articles.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Matteo Giomi, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. A unified framework for quantifying privacy risk in synthetic data. *arXiv preprint arXiv:2211.10459*, 2022.

Nicolas Grislain. Rag with differential privacy. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pages 847–852. IEEE, 2025.

Longzhu He, Peng Tang, Yuanhe Zhang, Pengpeng Zhou, and Sen Su. Mitigating privacy risks in retrieval-augmented generation via locally private entity perturbation. *Inf. Process. Manag.*, 62(5): 104150, 2025.

Xiao Hu, Eric Liu, Weizhou Wang, Xiangyu Guo, and David Lie. MARAGE: transferable multi-model adversarial attack for retrieval-augmented generation data extraction. *CoRR*, arXiv:2502.04360, 2025. doi: 10.48550/arXiv.2502.04360.

Palak Jain, Sofya Raskhodnikova, Satchit Sivakumar, and Adam D. Smith. The price of differential privacy under continual observation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 14654–14678. PMLR, 2023.

Changyue Jiang, Xudong Pan, Geng Hong, Chenfu Bao, and Min Yang. RAG-thief: Scalable extraction of private data from retrieval-augmented generation applications with agent-based attacks. *CoRR*, arXiv:2411.14110, 2024. doi: 10.48550/arXiv.2411.14110.

Nafiz Imtiaz Khan and Vladimir Filkov. Evidencebot: A privacy-preserving, customizable RAG-based tool for enhancing large language model interactions. In Leonardo Montecchi, Jingyue Li, Denys Poshyvanyk, and Dongmei Zhang, editors, *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering, FSE Companion 2025, Clarion Hotel Trondheim, Trondheim, Norway, June 23-28, 2025*, pages 1188–1192. ACM, 2025.

Tatsuki Koga, Ruihan Wu, Zhiyuan Zhang, and Kamalika Chaudhuri. Privacy-preserving retrieval augmented generation with differential privacy. *CoRR*, arXiv:2412.04697, 2024. doi: 10.48550/arXiv.2412.04697.

Mingrui Liu, Sixiao Zhang, and Cheng Long. Mask-based membership inference attacks for retrieval-augmented generation. In Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov, editors, *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, pages 2894–2907. ACM, 2025.

Christian Di Maio, Cristian Cosci, Marco Maggini, Valentina Poggioni, and Stefano Melacci. Pirates of the RAG: adaptively attacking llms to leak knowledge bases. *CoRR*, abs/2412.18295, 2024. doi: 10.48550/arXiv.2412.18295.

Qianren Mao, Qili Zhang, Hanwen Hao, Zhentao Han, Runhua Xu, Weifeng Jiang, Qi Hu, Zhijun Chen, Tyler Zhou, Bo Li, Yangqiu Song, Jin Dong, Jianxin Li, and Philip S. Yu. Privacy-preserving federated embedding learning for localized retrieval-augmented generation. *CoRR*, arXiv:2504.19101, 2025. doi: 10.48550/arXiv.2504.19101.

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*, pages 1216–1235, 2026.

Ali Naseh, Yuefeng Peng, Anshuman Suri, Harsh Chaudhari, Alina Oprea, and Amir Houmansadr. Riddle me this! stealthy membership inference for retrieval-augmented generation. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, page 1245–1259. Association for Computing Machinery, 2025. doi: 10.1145/3719027.3744840.

Yuefeng Peng, Junda Wang, Hong Yu, and Amir Houmansadr. Data extraction attacks in retrieval-augmented generation via backdoors. *CoRR*, arXiv:2411.01705, 2024. doi: 10.48550/arXiv.2411.01705.

Zhenting Qi, Hanlin Zhang, Eric P. Xing, Sham M. Kakade, and Himabindu Lakkaraju. Follow my instruction and spill the beans: Scalable data extraction from retrieval-augmented generation systems. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Bilal Rahib. fiqa-personal-finance-dataset. `https://huggingface.co/datasets/bilalRahib/fiqa-personal-finance-dataset`.

Ellery Sabado. Comprehensive-Medical-QA-Dataset. https://huggingface.co/datasets/elsabado/Comprehensive-Medical-QA-Dataset.

Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data - anonymisation groundhog day. In Kevin R. B. Butler and Kurt Thomas, editors, *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, pages 1451–1468. USENIX Association, 2022.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Maya Varma, Laurel Orr, Sen Wu, Megan Leszczynski, Xiao Ling, and Christopher Ré. Cross-domain data integration for named entity disambiguation in biomedical text. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4566–4575, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Baiqiang Wang, Qian Lou, Mengxin Zheng, and Dongfang Zhao. PIR-RAG: A system for private information retrieval in retrieval-augmented generation. *CoRR*, arXiv:2509.21325, 2025. doi: 10.48550/arXiv.2509.21325.

Yuhao Wang, Wenjie Qu, Shengfang Zhai, Yanze Jiang, Liu Zichen, Yue Liu, Yinpeng Dong, and Jiaheng Zhang. Silent leaks: Implicit knowledge extraction attack on rag systems. In *International Conference on Learning Representations*, volume 2026, pages 24150–24191, 2026.

Ruihan Wu, Erchi Wang, and Yu-Xiang Wang. Beyond per-question privacy: Multi-query differential privacy for RAG systems. In *NeurIPS 2025 Workshop on Reliable Machine Learning in the Wild*, 2025. OpenReview preprint.

Peiru Yang, Jinhua Yin, Haoran Zheng, Xueying Bai, Huili Wang, Yufei Sun, Xintian Li, Songwei Pei, Yongfeng Huang, and Tao Qi. Mrm: Black-box membership inference attacks against multimodal rag systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34295–34303, 2026.

Dixi Yao and Tian Li. Differentially private retrieval augmented generation with random projection. In *ICLR 2025 Workshop on Building Trust in LLMs and LLM Applications: From Guardrails to Explainability to Regulation*, 2025.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yiding Liu, Yue Xing, Han Xu, Jie Ren, Yi Chang, Shuaiqiang Wang, Dawei Yin, et al. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4505–4524, 2024.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Jie Ren, Tianqi Zheng, Hanqing Lu, Han Xu, Hui Liu, Yue Xing, and Jiliang Tang. Mitigating the privacy issues in retrieval-augmented generation (RAG) via pure synthetic data. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, page 24527–24558, 2025.

Fu Zichuan. HealthCareMagic-100k. https://huggingface.co/datasets/fzkuji/HealthCareMagic-100k.

A NOTATION

Table 1: Notation for answer-bearing retrieval and private synopsis fitting.

Symbol	Meaning
$D, D_{\mathrm{pub}}, \Omega$	Protected evidence chunks, public documents, domain policy.
$\mathcal{P}_{\mathrm{pub}}, \mathcal{A}_c$	All public chunks and eligible category representatives.
a, S_c	Public source passage and one sampled category representative.
$Q_c, m_c, c^*(q)$	Seed set, seed count, public query router.
$E, C(q,a), a_C$	Source selector, sentence identifiers, answer-bearing text.
$b(q,a)$	Public assessment of answer coverage.
ϕ, r, N_k	Encoder, joint query–representative retrieval score, evidence set.
$M(q,d,s), \chi(q,a,d)$	Private evidence-support score and applicability predicate.
$f_{q,a}, s_{q,a}, \tau, \mu$	Bounded support, feedback, target, margin.
$P_{s,c}^{(i)}, P_{q,c}^{(i)}$	Private synopsis and query fitting distributions.
g_i, h_i, t	Representative score, query hardness, fitting iterations.
$\alpha_{s,i}, \alpha_{q,i}$	Calibrated step sizes.
$\varepsilon_c, \varepsilon_v, \varepsilon_{\mathrm{life}}$	Category, version, and lifetime privacy allocations.
$\widehat{\ell}_{\mu,c}, \Gamma_c, R_c$	Private ramp loss, mean feedback, low-score rate.
$W(q,a), \xi_c(P)$	Independent validity label and high-score invalidity mass.
ρ, η	Population and synopsis-realization failure probabilities.
B, k, W	Online evidence budget, private retrieval size, monitor window.

B COMPLETE ALGORITHMS AND CONSTRUCTION

This section specifies public candidate construction, answer-bearing scoring, and private synopsis release. Algorithm 1 constructs a single synopsis inside the trusted curator, releasing one complete public passage per category. Appendix B.4 additionally describes an optional refresh controller in Algorithm 2.

Configuration and interfaces. Let Θ collect the domain policy Ω, encoder ϕ, nominal category count K, evidence-token budget B, private retrieval size k, retrieval weight λ, fitting length t, feedback target τ, and margin μ. It also records source and metadata rules, segmentation settings, model and tokenizer revisions, prompts, decoding settings, length limits, and tie-breaking rules. We require $K, k, t \geq 1$, $B > 0$, $0 < \lambda < 1$, and $0 < \mu \leq \tau \leq 1$. These choices and the public candidate space are fixed before accessing the private snapshot for the release. A refresh may derive its configuration from public inputs and earlier released outputs, as specified below.

The private snapshot D is a multiset of indexed evidence windows with immutable record identifiers and applicability metadata. The public corpus D_{pub} contains approved source documents with immutable source identifiers. The seed workload Q is an indexed list: repeated query texts remain separate workload entries. An artifact $\mathcal{R}$ contains sampled public passages, their source metadata, and a public router. The fixed empty artifact $\mathcal{R}_\perp$ has no passages and returns $\perp$ on routing and lookup.

B.1 PUBLIC CONSTRUCTION AND ANSWER VIEWS

Source-linked passage construction. $\mathrm{SentenceChunks}$ processes every approved public document in source order using the declared sentence splitter and target chunk length. Boundaries fall between complete sentences; an overlong sentence remains intact at this stage. Each passage retains its original text, source identifier, sentence identifiers, offsets, and approved public metadata. Segmentation depends on the public documents and configuration alone.

$\mathrm{Eligible}$ applies the approved-source and applicability policies. Fits checks the length of the complete passage together with its citation delimiters and serialized metadata under the generator's tokenizer against B. Passages that exceed this budget are excluded during eligibility filtering. Exact duplicate passages with identical applicability metadata may be represented by one canonical public identifier under a fixed deduplication rule. Passages with different applicability metadata remain distinct.

Algorithm 1: BUILDSYNOPSIS: Answer-bearing private synopsis construction

Require: Private snapshot D, public corpus D_{pub}, seed queries Q, public configuration Θ, reserved release budget ε
Ensure: Public artifact $\mathcal{R} = (\{S_c\}_{c\in\mathcal{C}}, \mathrm{Route})$

// Phase I: Public passage pools
$\mathcal{P}_{\mathrm{pub}} \leftarrow \mathrm{SentenceChunks}(D_{\mathrm{pub}}; \Theta)$
$\mathcal{A} \leftarrow \{a \in \mathcal{P}_{\mathrm{pub}} : \mathrm{Eligible}(a; \Omega) \wedge \mathrm{Fits}(a; B)\}$
$(\{\mathcal{A}_c\}_{c\in\mathcal{C}}, \mathrm{Route}) \leftarrow \mathrm{PublicPartition}(\mathcal{A}; \phi, K, \Theta)$
if $\mathcal{C} = \varnothing$ **then**
  return $\mathcal{R}_\perp$
end if
// Phase II: Private scoring, fitting, and sampling
for each $c \in \mathcal{C}$ **do**
  $Q_c \leftarrow [q \in Q : \mathrm{Route}(q) = c]$; $\varepsilon_c \leftarrow \varepsilon/|\mathcal{C}|$
  if $Q_c = \varnothing$ **then**
    $S_c \leftarrow \mathrm{FirstByPublicID}(\mathcal{A}_c)$
  else if $\varepsilon_c = 0$ **then**
    $S_c \sim \mathrm{Unif}(\mathcal{A}_c)$
  else
    for each $(q, a) \in Q_c \times \mathcal{A}_c$ **do**
      $(C(q,a), b(q,a)) \leftarrow \mathrm{FLANSelectAssess}(q, a)$; $f_{q,a} \leftarrow 0$
      if $C(q,a) \neq \varnothing$ **and** $b(q,a) > 0$ **then**
        $a_C \leftarrow \mathrm{ConcatSourceSentences}(a, C(q,a))$
        $D_{q,a} \leftarrow \{d \in D : \chi(q,a,d) = 1\}$
        $r_d \leftarrow \lambda\,\mathrm{sim}_{01}(\phi(q), \phi(d)) + (1-\lambda)\,\mathrm{sim}_{01}(\phi(a_C), \phi(d))$, $d \in D_{q,a}$
        $N \leftarrow \mathrm{TopK}(D_{q,a}, \{r_d\}, k)$
        if $N \neq \varnothing$ **then**
          $u_d \leftarrow \min_{s\in C(q,a)} \mathrm{MiniCheck}(d, s)$, $d \in N$
          $f_{q,a} \leftarrow b(q,a) \max_{d\in N}\{\chi(q,a,d) u_d\}$
        end if
      end if
      $s_{q,a} \leftarrow \min\{1, \max\{-1,\ 2(f_{q,a} - (\tau - \mu))/\mu - 1\}\}$
    end for
    $P^{(1)}_{s,c} \leftarrow \mathrm{Unif}(\mathcal{A}_c)$; $P^{(1)}_{q,c} \leftarrow \mathrm{Unif}(Q_c)$
    $\alpha_s \leftarrow \varepsilon_c/(4t)$; $\alpha_q \leftarrow \alpha_s$
    for $i = 1, \ldots, t$ **do**
      $g_i(a) \leftarrow \sum_{q\in Q_c} P^{(i)}_{q,c}(q) s_{q,a}$, $a \in \mathcal{A}_c$
      $h_i(q) \leftarrow -\sum_{a\in\mathcal{A}_c} P^{(i)}_{s,c}(a) s_{q,a}$, $q \in Q_c$
      $\widetilde{P}_s(a) \leftarrow P^{(i)}_{s,c}(a) \exp(\alpha_s g_i(a))$, $a \in \mathcal{A}_c$
      $\widetilde{P}_q(q) \leftarrow P^{(i)}_{q,c}(q) \exp(\alpha_q h_i(q))$, $q \in Q_c$
      $P^{(i+1)}_{s,c} \leftarrow \widetilde{P}_s / \sum_{a\in\mathcal{A}_c} \widetilde{P}_s(a)$
      $P^{(i+1)}_{q,c} \leftarrow \widetilde{P}_q / \sum_{q\in Q_c} \widetilde{P}_q(q)$
    end for
    $S_c \sim P^{(t+1)}_{s,c}$ {fresh secret sampling coins}
  end if
end for
return $\mathcal{R} \leftarrow (\{\mathrm{PublicPassageWithMetadata}(S_c)\}_{c\in\mathcal{C}}, \mathrm{Route})$

Category pools and routing. PublicPartition groups eligible passages using their public embeddings or a declared public domain partition. Embedding-based construction uses at most K clusters, with fixed initialization, an iteration cap, and immutable-ID tie-breaking. Empty clusters are removed, yielding a category index set $\mathcal{C}$ and disjoint nonempty pools $\{\mathcal{A}_c\}_{c\in\mathcal{C}}$. Any per-category candidate cap is applied using a declared public selection rule before private scoring. Define

$$z_c = \frac{1}{|\mathcal{A}_c|}\sum_{a\in\mathcal{A}_c}\phi(a), \qquad \nu_c = \begin{cases} z_c/\|z_c\|_2, & \|z_c\|_2 > 0, \\ 0, & \text{otherwise}, \end{cases} \qquad \mathrm{Route}(q) = \arg\max_{c\in\mathcal{C}} \cos_0(\phi(q), \nu_c).$$

The router uses a fixed category ordering to resolve ties, and returns $\perp$ if $\mathcal{C}$ is empty. Its cosine convention is defined in Appendix B.2. Seed entries are routed into $Q_c = [q \in Q : \mathrm{Route}(q) = c]$. The resulting pools and router are independent of the private evidence. For the population

utility theorem, the pools, router, and complete scoring configuration must additionally be fixed independently of the seed sample used to estimate risk.

Answer-bearing sentence selection. For each (q, a), FLANSelectAssess uses the frozen `google/flan-t5-xl` checkpoint (Chung et al., 2024) in evaluation mode with greedy decoding and a 128-token output cap. Sentences in a are numbered in source order. The fixed prompt is:

> Given the question and numbered passage, select the sentence IDs that supply an answer. Include sentences containing conditions, exceptions, or referents needed to interpret that answer. Assess the selected content's coverage of the question: 0 = absent, 1 = partial, 2 = complete. Return one line: IDs: [comma-separated IDs]; coverage: [0, 1, or 2]. Use an empty ID list and coverage 0 for absent answer content.
> Question: *query text*
> Passage: *numbered source sentences*

The parser validates the returned identifiers and coverage label, removes repeated sentence identifiers, and restores source order. The valid identifiers define $C(q, a)$; the coverage label divided by two defines $b(q, a) \in \{0, 1/2, 1\}$. ConcatSourceSentences concatenates the selected source sentences without changing their wording to form a_C. Malformed output, invalid identifiers, an invalid coverage label, an empty selection, or coverage zero gives $(C(q, a), b(q, a)) = (\varnothing, 0)$. A formatted selector input exceeding the declared 512-token budget receives the same outcome. The full source passage remains the release candidate, while a_C is its query-specific scoring view.

B.2 Private Evidence Contract

Applicability and private retrieval. Each private evidence window carries its parent-document identifier and the metadata required by the domain policy. A fixed policy table maps the query and candidate's public scope labels to required private metadata fields and allowed values or ranges. The predicate $\chi(q, a, d)$ is one exactly when the required entity, population, period, and scope checks pass; missing required metadata gives zero. Fields outside the declared requirements impose no additional restriction. The eligible evidence set is $D_{q,a} = \{d \in D : \chi(q, a, d) = 1\}$.

For a valid answer view, the standard dense retriever ranks eligible windows by

$$r(q, a, d) = \lambda \, \mathrm{sim}_{01}(\phi(q), \phi(d)) + (1 - \lambda) \, \mathrm{sim}_{01}(\phi(a_C), \phi(d)), \qquad 0 < \lambda < 1, \tag{16}$$

where $\mathrm{sim}_{01}(x, y) = (1 + \cos_0(x, y))/2$. For nonzero vectors, $\cos_0$ is cosine similarity clipped to $[-1, 1]$; if either vector is zero it is -1. Malformed encoder outputs are represented by zero vectors. The encoder checkpoint and input-length policy are fixed in Θ and shared by clustering, routing, private ranking, and query monitoring for that version.

TopK returns the highest-ranked $\min\{k, |D_{q,a}|\}$ windows using exact ranking and immutable-ID tie-breaking. The default configuration uses $\lambda = 1/2$ and $k = 3$. Retrieval is performed for each seed–candidate pair, since a_C depends on the public candidate. An empty eligible set yields no retrieved evidence and fidelity zero.

Sentence support and fidelity. We use `lytang/MiniCheck-Flan-T5-Large` (Tang et al., 2024) for support assessment. Its document input consists of one private evidence window and policy-approved metadata serialized in a fixed field order; its claim input is one unchanged source sentence $s \in C(q, a)$. The supported-class score returned by the authors' `score(docs, claims)` interface, clipped to $[0, 1]$, defines $M(q, d, s)$. The pseudocode's $\mathrm{MiniCheck}(d, s)$ denotes this configured interface. An invalid score or a formatted pair exceeding the declared 512-token checker budget receives support zero.

Within each retrieved window, define $u_d = \min_{s \in C(q,a)} M(q, d, s)$. The fidelity score is

$$f_{q,a}(D) = \begin{cases} b(q, a) \max\limits_{d \in N_k(D; q, a)} \{\chi(q, a, d) u_d\}, & C(q, a) \neq \varnothing, \; N_k(D; q, a) \neq \varnothing, \\ 0, & \text{otherwise.} \end{cases}$$

Thus every selected sentence must receive support from the same retrieved window, and the strongest applicable window determines the passage's support. Combined with the completeness assessment, this gives $f_{q,a} \in [0, 1]$. The score estimates answer-bearing support under the declared scope and retrieved evidence; its semantic accuracy depends on the selector, assessor, metadata, and checker.

B.3 Fitting, Sampling, and the Release Interface

The curator computes the feedback table once for each fixed snapshot, seed list, and candidate pool:

$$s_{q,a} = \min\{1, \max\{-1,\ 2(f_{q,a} - (\tau - \mu))/\mu - 1\}\}.$$

This map assigns -1 to fidelity at or below $\tau - \mu$, $+1$ to fidelity at or above τ, and interpolates linearly between them. For each category with seeds and positive budget, both fitting distributions start uniformly. Each round computes g_i and h_i from the same iteration-i distributions before forming either updated distribution. The normalization in Algorithm 1 therefore implements simultaneous updates.

The release budget is divided equally across the nonempty public category pools: $\varepsilon_c = \varepsilon/|\mathcal{C}|$. Since every candidate update lies in $[-1, 1]$, the final representative sample satisfies the privacy bound $4\sum_i \alpha_{s,i}$. We therefore use

$$\alpha_s = \alpha_q = \frac{\varepsilon_c}{4t} = \frac{\varepsilon}{4|\mathcal{C}|t}.$$

After t rounds, the curator draws one representative from each fitted candidate distribution with fresh secret randomness. A category without seed queries uses its first candidate under the fixed public-ID ordering. A seeded category with zero budget samples uniformly from its public pool. These public-only branches require no private scoring; their reserved allocation can remain conservatively charged in the ledger.

The artifact contains the sampled complete passages, their public source identifiers, offsets and metadata, and the router. Retrieved private windows, support scores, feedback tables, and fitting distributions remain inside the curator. An artifact is constructed internally and becomes observable only at its assigned publication slot. The privacy analysis assumes exact updates and exact categorical sampling; an implementation using finite precision requires a sampling and numerical analysis consistent with that guarantee.

B.4 Optional Offline Refresh and Online Serving

The primary method constructs and reuses a single synopsis. This optional extension reconstructs the synopsis when public resources, service requirements, or the protected evidence snapshot change. Each additional release receives a separate privacy allocation under a predeclared lifetime budget.

Updating public candidate pools or routing may improve coverage of a changed workload. Refitting over unchanged pools remains subject to Proposition 2 and does not guarantee improved accuracy. The controller below specifies when reconstruction occurs and how releases are accounted for; its utility under distribution shift is not evaluated in the fixed-synopsis experiments.

Serving and controller state. Let Θ be the configuration of the installed artifact and Θ_u the current announced public configuration at time u. A query is admissible when it satisfies the current public scope and query-length rules. For an admissible query, the installed router chooses one category, and Lookup returns its released passage. $\mathrm{Allowed}$ checks current source eligibility and serving constraints. The generator receives the complete allowed passage with its public metadata and the fixed grounding prompt. Missing or disallowed context yields the fixed insufficient-evidence response $o_\perp$; inadmissible queries receive the public refusal response.

The controller maintains the installed artifact $\mathcal{R}$, its seed list Q, snapshot time s_{cur}, publication time p_{cur}, and a buffer $\mathcal{B}$ containing at most the last $W \geq 1$ accepted queries since publication. $J = \perp$ denotes an idle builder. A pending record J contains the assigned publication slot $J.p$, launch snapshot time $J.s$, new seed list $J.Q$, public configuration $J.\Theta$, and private build result $J.\mathrm{result}$. The next unused version index is v, and L is the total reserved budget including initialization.

Public monitoring signals. Using the installed version's encoder, define

$$\sigma(q, Q; \Theta) = \max_{q' \in Q} \mathrm{sim}_{01}(\phi_\Theta(q), \phi_\Theta(q')),$$
$$\rho(q, \mathcal{R}; \Theta, \Theta_u) = \mathrm{sim}_{01}(\phi_\Theta(q), \phi_\Theta(S_{\mathrm{Route}_{\mathcal{R}}(q)})).$$

Set $\sigma = 0$ for an empty seed list and $\rho = 0$ for an absent or currently disallowed representative. At a checkpoint with a full buffer, the drift fraction and context-gap fraction are

$$d = \frac{1}{W} \sum_{q \in \mathcal{B}} \mathbf{1}\{\sigma(q, Q; \Theta) < \gamma_Q\}, \qquad g = \frac{1}{W} \sum_{q \in \mathcal{B}} \mathbf{1}\{\rho(q, \mathcal{R}; \Theta, \Theta_u) < \gamma_S\}.$$

Both fractions are zero until the buffer is full. Fix relevance thresholds $\gamma_Q, \gamma_S \in (0, 1]$, trigger fractions $\theta_Q, \theta_S \in (0, 1]$, and intervals $\Delta_{\text{age}} \geq \Delta_{\min} > 0$. A refresh request occurs when either fraction exceeds its trigger, the announced scope/source revision Ω_u differs from the installed revision Ω, or the snapshot age reaches Δ_{age}.

Reservation, reconstruction, and publication. Initialization and refreshes have finitely many preallocated caps $\varepsilon_0, \ldots, \varepsilon_V$ satisfying $\sum_{v=0}^{V} \varepsilon_v \leq \varepsilon_{\text{life}}$. A requested refresh launches only when the builder is idle, the cooldown since publication has elapsed, and an unused cap remains. The controller reserves that cap irrevocably before private access and advances the next-slot index. This ensures that every launched job is charged once.

$\mathrm{PublicRebuild}(h_u)$ produces the new approved corpus and configuration from the permitted history h_u, which consists of public inputs and preceding released outputs. $\mathrm{Anchors}$ supplies admissible public anchor questions; these are concatenated with the buffered questions and filtered through the new public query policy. Repeated questions retain their indexed multiplicity. The new pools and router determine the new category assignments. The private snapshot is captured at launch, and Algorithm 1 recomputes the feedback and fits from uniform distributions.

The public schedule maps each launch opportunity u to a later publication slot $p(u)$, with $p_0 > s_0$ for initialization. The event stream contains query arrivals, public checkpoints, announced-policy changes, and scheduled publication events; publication events occur even if no query arrives at that time. Events use public tie-breaking, and an assigned publication is processed before serving a query at the same time. Under the ideal controller, the pending build is available at that slot, and publication atomically replaces the artifact and its configuration and clears the monitoring buffer. The previous artifact remains active during construction and after budget exhaustion, subject to current public eligibility rules. Observable publication, retry, and acceptance decisions follow this public schedule and permitted history; implementation failures or timing channels require additional treatment if they become observable.

Here $\|$ denotes list concatenation, Last_W retains the last W entries, and $\mathrm{AdmissibleFilter}$ preserves the order of entries satisfying the current query policy. $\mathrm{AnnouncedConfig}(h_u)$ reads the current public policy, and Ω_u is its scope/source revision. The asynchronous assignment starts one internal build using the captured snapshot; its result is read only at the assigned publication slot. Reusing already released passages during serving is post-processing, whereas each fresh sampled version is charged to the lifetime ledger.

B.5 Defaults and Computational Cost

The reported default configuration uses $K = 10$, $t = 2$, $\tau = 0.7$, $\mu = 0.4$, $k = 3$, and $\lambda = 0.5$, with five public seed entries and 32 public candidate chunks per category. Thus a fully populated default release scores $10 \times 5 \times 32 = 1600$ seed–candidate pairs. Each pair with a valid answer view requires one private retrieval and at most $k|C(q, a)|$ sentence-support evaluations. Public passage embeddings and answer views may be cached when their inputs and configuration are unchanged.

Writing $m_c = |Q_c|$ and $N_c = |\mathcal{A}_c|$, a dense feedback table uses $O(\sum_c m_c N_c)$ space, and the fitting stage requires $O(t \sum_c m_c N_c)$ arithmetic operations after scoring. These counts exclude encoder, retrieval, and model-inference costs. The published artifact stores one passage per nonempty category, while each online query uses only its routed passage. A refresh recomputes private feedback for the new snapshot and workload and incurs a separately reserved release cost.

C Proof of the Sampled-Release Privacy Guarantee

Public configuration and bounded updates. Fix a category's public candidate pool, indexed seed list, initialization, and nonnegative step-size sequences before the release's private access. For a refresh, condition on the same preceding observable history. The initial candidate distribution p_0

Algorithm 2: Optional offline refresh with lifetime privacy accounting

Require: Private snapshot interface $D(\cdot)$, initial public corpus $D_{\mathrm{pub},0}$, configuration Θ_0, seed list Q_0, snapshot time s_0, publication slot p_0; generator G; version caps $\{\varepsilon_v\}_{v=0}^{V}$ with $\sum_{v=0}^{V}\varepsilon_v \le \varepsilon_{\mathrm{life}}$; public checkpoints and slots $p(u)$; $W, \gamma_Q, \gamma_S, \theta_Q, \theta_S, \Delta_{\mathrm{age}} \ge \Delta_{\min} > 0$

Ensure: Released synopsis versions and online responses

$L \leftarrow \varepsilon_0$; $v \leftarrow 1$; $J \leftarrow \perp$; $\mathcal{B} \leftarrow [\,]$
$\mathcal{R} \leftarrow \textsc{BuildSynopsis}(D(s_0), D_{\mathrm{pub},0}, Q_0, \Theta_0, \varepsilon_0)$
$\mathrm{PublishAt}(p_0, \mathcal{R})$; $(Q, \Theta, s_{\mathrm{cur}}, p_{\mathrm{cur}}) \leftarrow (Q_0, \Theta_0, s_0, p_0)$
for each event e at time $u \ge p_0$ **do**
  $\Theta_u \leftarrow \mathrm{AnnouncedConfig}(h_u)$
  if $J \neq \perp$ **and** $u = J.p$ **then**
    $\mathcal{R} \leftarrow J.\mathrm{result}$; $\mathrm{Publish}(\mathcal{R})$
    $(Q, \Theta, s_{\mathrm{cur}}, p_{\mathrm{cur}}) \leftarrow (J.Q, J.\Theta, J.s, u)$; $\mathcal{B} \leftarrow [\,]$; $J \leftarrow \perp$
  end if
  if $e = \mathrm{Query}(q)$ **then**
    if $\neg\,\mathrm{Admissible}(q; \Theta_u)$ **then**
      $\mathrm{Output}(\mathrm{PublicRefusal}(q))$
    else
      $c \leftarrow \mathrm{Route}_{\mathcal{R}}(q)$; $a \leftarrow \mathrm{Lookup}(\mathcal{R}, c)$
      if $a = \perp$ **or** $\neg\,\mathrm{Allowed}(a; \Theta_u)$ **then**
        $\mathrm{Output}(o_\perp)$
      else
        $\mathrm{Output}(G(q, a))$
      end if
      $\mathcal{B} \leftarrow \mathrm{Last}_W(\mathcal{B} \,\|\, [q])$
    end if
  else if $e = \mathrm{Checkpoint}$ **then**
    $(d, g) \leftarrow (0, 0)$
    if $|\mathcal{B}| = W$ **then**
      $d \leftarrow W^{-1} \sum_{q \in \mathcal{B}} \mathbf{1}\{\sigma(q, Q; \Theta) < \gamma_Q\}$
      $g \leftarrow W^{-1} \sum_{q \in \mathcal{B}} \mathbf{1}\{\rho(q, \mathcal{R}; \Theta, \Theta_u) < \gamma_S\}$
    end if
    $z \leftarrow (d \ge \theta_Q) \vee (g \ge \theta_S) \vee (\Omega_u \neq \Omega) \vee (u - s_{\mathrm{cur}} \ge \Delta_{\mathrm{age}})$
    if z **and** $J = \perp$ **and** $u - p_{\mathrm{cur}} \ge \Delta_{\min}$ **and** $v \le V$ **then**
      $\varepsilon_{\mathrm{job}} \leftarrow \varepsilon_v$; $L \leftarrow L + \varepsilon_{\mathrm{job}}$; $v \leftarrow v + 1$ {reserve and retire the slot}
      $(D^{+}_{\mathrm{pub}}, \Theta^{+}) \leftarrow \mathrm{PublicRebuild}(h_u)$
      $Q^{+} \leftarrow \mathrm{AdmissibleFilter}(\mathrm{Anchors}(\Theta^{+}) \,\|\, \mathcal{B}; \Theta^{+})$
      $D^{+} \leftarrow D(u)$; $J \leftarrow (p : p(u),\ s : u,\ Q : Q^{+},\ \Theta : \Theta^{+},\ \mathrm{result} : \perp)$
      $J.\mathrm{result} \leftarrow$ **async** $\textsc{BuildSynopsis}(D^{+}, D^{+}_{\mathrm{pub}}, Q^{+}, \Theta^{+}, \varepsilon_{\mathrm{job}})$
    end if
  end if
end for

and candidate support are then identical in neighboring worlds. Public construction randomness can be conditioned on and subsequently integrated out because its conditional law is the same in both worlds.

The fidelity score lies in $[0, 1]$ and the clipped feedback satisfies $s_{q,a}(D) \in [-1, 1]$ for every corpus D. Since $P^{(i),D}_{q,c}$ is a probability distribution,

$$g_i^D(a) = \sum_{q \in Q_c} P^{(i),D}_{q,c}(q)\, s_{q,a}(D) \in [-1, 1].$$

Consequently, for every neighboring pair D, D',

$$|g_i^D(a) - g_i^{D'}(a)| \le 2. \tag{17}$$

This bound includes the private-data dependence of the query weights. It holds at every round regardless of how the two query distributions differ. The query-step sequence is fixed as part of the release configuration; its effect on privacy is controlled through the bounded candidate update g_i.

Final sampled representative. Unrolling the candidate updates gives

$$P_s^{(t+1),D}(a) = \frac{p_0(a)e^{G_D(a)}}{Z_s(D)}, \qquad G_D(a) = \sum_{i=1}^{t} \alpha_{s,i} g_i^D(a), \qquad Z_s(D) = \sum_{b \in \mathcal{A}_c} p_0(b) e^{G_D(b)}. \tag{18}$$

Let $A_s = \sum_{i=1}^{t} \alpha_{s,i}$. Equation (17) implies

$$|G_D(a) - G_{D'}(a)| \le 2A_s.$$

Multiplication by the common prior and summation over candidates therefore give

$$e^{-2A_s} \le \frac{Z_s(D')}{Z_s(D)} \le e^{2A_s}.$$

Uniform initialization has full support, so each pointwise likelihood ratio is defined and obeys

$$\frac{P_s^{(t+1),D}(a)}{P_s^{(t+1),D'}(a)} = e^{G_D(a) - G_{D'}(a)} \frac{Z_s(D')}{Z_s(D)} \le e^{4A_s}. \tag{19}$$

For any event $O \subseteq \mathcal{A}_c$, summing this inequality over O gives

$$\Pr[S_c(D) \in O] \le e^{4A_s} \Pr[S_c(D') \in O].$$

Exchanging the two corpora gives the reverse inequality. Thus one final representative sample is $4A_s$-DP, establishing the sufficient calibration

$$4 \sum_{i=1}^{t} \alpha_{s,i} \le \varepsilon_c.$$

For the implemented schedule $\alpha_{s,i} = \alpha_{q,i} = \varepsilon_c/(4t)$, the certified representative bound is $4t\alpha_s = \varepsilon_c$. The equality of the two step sizes is the algorithm's fitting choice; the proof analyzes the actual representative release directly. It covers the influence of both internal fitting distributions without introducing an additional sampled query output. Mapping the sampled public identifier to its complete source passage and public provenance is post-processing.

Public-only branches and category composition. The empty artifact, the fixed representative of a seedless category, and the uniform sample of a zero-budget category have distributions independent of D under the conditioned public configuration, and therefore incur zero privacy loss. Whether these branches apply is determined by the public pools, seed lists, and allocated budgets. In the fitted branches, every category may use the same private corpus, so sequential composition gives at most $\sum_{c \in \mathcal{C}} \varepsilon_c \le \varepsilon$ for the complete released artifact (Dwork and Roth, 2014). Fresh independent sampling coins satisfy the required conditional release bounds; shared public construction choices can be integrated out after conditioning.

Adaptive serving and further versions. Fix the same external query and announcement strategies, auxiliary information, and data-independent randomization in neighboring worlds. For a fixed released artifact and any finite horizon, routing, generation, public refusals, context fallback, and buffering form a randomized channel of that artifact and the external inputs. Post-processing therefore gives the same privacy bound for the joint synopsis-and-answer transcript. This statement protects the private evidence corpus; user queries are the service inputs under this adjacency relation.

For a subsequent version, conditioning on the preceding transcript fixes the public controller decisions, new construction inputs, reserved cap, and uniform initialization. The same release proof then applies conditionally to that version. Composition over the reserved version caps yields a bound of $\sum_{v=0}^{V} \varepsilon_v \le \varepsilon_{\text{life}}$ under the neighboring-history relation specified for the refresh mechanism. Monitoring and assigned publication events use permitted history and add no separate private-data access.

Scope of the release guarantee. The guarantee covers the sampled public artifact and its post-processing. A fresh sample from an unreleased fitted distribution is a new release and requires accounting. Exact fitted weights, raw support diagnostics, private-score-based acceptance decisions, and observable private-dependent completion behavior are outside the stated output interface. The proof relies on bounded scores and controlled randomization; factual reliability of the checker and correctness of the generated answer are separate utility questions.

D Proofs and Interpretation of Answer-Support Bounds

Finite-workload average. Define $\ell_\mu(q,a) = (1 - s_{q,a})/2 \in [0,1]$. Whenever $f_{q,a} \le \tau - \mu$, feedback is -1, so

$$\mathbf{1}\{f_{q,a} \le \tau - \mu\} \le \ell_\mu(q,a).$$

Averaging over independent draws $q \sim \mathrm{Unif}(Q_c)$ and $S_c \sim P$ proves Eq. (11). Linearity gives $\widehat{\ell}_{\mu,c}(P) = (1 - \Gamma_c(P))/2$. No convergence claim about the updates is used.

Population statement and data-adaptive selection. Condition on the private corpus, finite public representative pool, encoder, answer selector, answer assessor, support checker, and router. Under the i.i.d. seed assumption, for each fixed representative define $\ell_c(a) = \mathbb{E}_{q\sim\Pi_c}\ell_\mu(q,a)$ and its empirical mean $\widehat{\ell}_c(a)$. A Hoeffding bound and union bound give

$$\Pr\left[\exists a \in \mathcal{A}_c : \ell_c(a) - \widehat{\ell}_c(a) > r\right] \le 2|\mathcal{A}_c| e^{-2m_c r^2}.$$

For $r = \sqrt{\log(2|\mathcal{A}_c|/\rho)/(2m_c)}$, the complementary event holds with probability at least $1-\rho$. On this event it holds simultaneously for every representative, hence for every distribution $\widetilde{P}$ on the pool:

$$\sum_a \widetilde{P}(a)\ell_c(a) \le \sum_a \widetilde{P}(a)\widehat{\ell}_c(a) + r.$$

This includes the final fitted P, despite its dependence on the seed workload and private feedback. Pointwise domination of failure by ramp loss proves Eq. (12). The argument requires the pool, router, and full scoring contract to be fixed independently of the seed sample under consideration; selecting a new pool using that same sample would require another generalization argument.

A one-time realized release. Equation (11) states $\mathbb{E}_{S_c\sim P} R_c(S_c) \le \widehat{\ell}_{\mu,c}(P)$. If the right side is zero, nonnegativity gives $R_c(S_c) = 0$ almost surely. Otherwise Markov's inequality bounds the probability that $R_c(S_c) > \widehat{\ell}_{\mu,c}(P)/\eta$ by η. Since $R_c \le 1$, Eq. (13) follows. The population version first conditions on the good seed event and then applies the same reasoning over the final representative draw. A union bound gives joint failure at most $\rho + \eta$.

These guarantees can be vacuous and do not make the private risks publicly releasable. They also do not imply that the final iterate is as good as an average iterate or an optimal representative. A no-regret theorem for an average strategy would need its own assumptions and would not automatically apply to the final distribution released here. The analyzed representative is the complete source passage supplied online. The selected sentence set is only a scoring view. Correctness of omitted qualifiers or other statements in the passage is not established by a high score.

Conditional semantic evidence bound. For an independently specified validity label W, an invalid representative either has $f_{q,a} \le \tau - \mu$ or has $f_{q,a} > \tau - \mu$ and $W(q,a) = 0$. These events partition invalidity according to the score threshold, with the low-score event possibly also containing valid representatives. Thus

$$\mathbf{1}\{W(q,a) = 0\} \le \mathbf{1}\{f_{q,a} \le \tau - \mu\} + \mathbf{1}\{f_{q,a} > \tau - \mu,\ W(q,a) = 0\}.$$

Averaging and applying Eq. (11) proves Proposition 1. The error term is weighted by the actual fitted distribution, so its definition remains valid for data-adaptive selection; it is not automatically known or small. An independent annotation protocol must assess whether the public answer is sufficient, correct, and applicable, including conditions omitted by E. Neither the support model's own scores nor their privacy guarantee supply these labels.

Let $U_c(a) = \Pr_{q\sim\mathrm{Unif}(Q_c)}[W(q,a) = 0]$. Since $\mathbb{E}_{S_c\sim P} U_c(S_c) \le \widehat{\ell}_{\mu,c}(P) + \xi_c(P)$, the same Markov argument gives

$$\Pr_{S_c\sim P}\left[U_c(S_c) \le \min\{1, (\widehat{\ell}_{\mu,c}(P) + \xi_c(P))/\eta\}\right] \ge 1 - \eta.$$

This is an evidence statement, not an answer-generation guarantee. The router sends each seed to its fitted category and serving supplies the complete sampled passage under the fixed public configuration. A later source-policy change can remove that passage; privacy is preserved but the earlier utility interpretation need not hold. Publishing private risks or semantic-error estimates requires separate accounting.

E PRIVACY ACCOUNTING FOR OPTIONAL REFRESHES

Initialization and subsequent releases receive preallocated budgets satisfying

$$\sum_{v=0}^{V} \sum_{c \in \mathcal{C}_v} \varepsilon_{v,c} \leq \sum_{v=0}^{V} \varepsilon_v \leq \varepsilon_{\text{life}}, \tag{20}$$

where $\mathcal{C}_v$ denotes the category set of version v.

Theorem 3 (Lifetime privacy)**.** *Let neighboring histories differ by insertion or removal of one protected chunk trajectory at all public snapshot opportunities, with every other chunk identical. Fix the same external input strategies in both worlds, allowing adaptation only to preceding released outputs and data-independent randomness. Under the controller in Appendix B.4 and the budget constraint in Eq.* (20)*, the transcript of synopsis versions, monitoring and reservation decisions, publication events, and answers is pure $\varepsilon_{\text{life}}$-DP at every finite horizon.*

Histories and permitted control decisions. Index the private corpus trajectory by all public snapshot opportunities, including initialization. Neighboring histories omit one complete chunk trajectory wherever it occurs and agree on all other chunks at every opportunity. Fix the same external query/public-announcement strategy with the same auxiliary information in both worlds. Queries and announcements may depend on earlier released outputs and data-independent randomness, but not on new unaccounted private diagnostics. Domain changes do not change the protected-unit definition or reset its lifetime ledger.

Let h be the preceding observable history, including released representatives, queries, monitoring decisions, and version events. Conditional on a fixed h, monitoring signals, cooldown state, selected checkpoint, allocation, public pool, and reconstructed seed workload have the same law in the two worlds. Public construction randomness may additionally be conditioned on. Both fitting distributions start from the same uniform weights. Therefore the next version's conditional release kernel satisfies Theorem 1, with total cost at most its reserved cap ε_v.

Monitor-only and serving events have identical conditional laws given their permitted inputs, so contribute no additional privacy factor. At a finite horizon the chain rule for conditional kernels gives the bound

$$\exp\left(\varepsilon_0 + \sum_{v \in \mathcal{R}(h)} \varepsilon_v\right) \leq \exp\left(\sum_{v=0}^{V} \varepsilon_v\right) \leq e^{\varepsilon_{\text{life}}}, \tag{21}$$

where $\mathcal{R}(h)$ contains refresh slots consumed on that history. Empty public branches and unused allocations have zero actual cost; charging reserved caps is conservative. Integrating the conditional bounds proves Theorem 3. The observable interface includes skipped checkpoints, refusals, and assigned publication slots. Private-dependent runtime or raw-quality-based retries cannot be added to this interface without further analysis.

Why initialization matters. A prior sampled representative is a DP output. Its underlying final fitting distribution is not. If that unreleased state were used as the next initial distribution, conditioning on the same observable history would not make the initial weights equal across neighboring corpora. Equation (18) would then have a data-dependent initial factor, and the stated per-version proof would no longer apply. Keeping that state local does not remove its influence on subsequent releases. The controller therefore refits from uniform public weights. A strictly positive prior computed only from earlier released representatives could also be analyzed conditional on history, but is not needed by the specified controller.

Utility after refresh. Conditional on a preceding history and the new snapshot, each refit satisfies the finite-workload average and realization bounds for its own newly constructed seed workload. For simultaneous realization guarantees over finitely many slots/categories, preallocate failure levels whose sum is at most a desired overall failure level and apply conditional Markov bounds and a union bound. The population result applies only if fresh seeds independently satisfy its i.i.d. assumption conditional on history. Queries selected by drift or low confidence generally do not. No monotone improvement across versions or continual factual-freshness theorem is asserted.

F Extension to Document-Level Privacy

Corollary 2 (Document-level privacy). *Let $T(X)$ be a fixed indexing rule mapping a source-document corpus X to evidence chunks. Under the public-configuration and initialization assumptions of Theorem 1, one sampled category representative is ε_c-DP for neighboring source corpora that differ by one whole document whenever*

$$4\sum_{i=1}^{t}\alpha_{s,i} \le \varepsilon_c.$$

Category releases compose to $\sum_c \varepsilon_c$ under this adjacency relation.

Proof. For every source corpus X, the bounded feedback implies

$$g_i^{T(X)}(a) = \sum_{q\in Q_c} P_{q,c}^{(i),T(X)}(q)\, s_{q,a}(T(X)) \in [-1,1].$$

Consequently, for neighboring source corpora X, X',

$$\left| g_i^{T(X)}(a) - g_i^{T(X')}(a) \right| \le 2,$$

regardless of how many indexed chunks the changed document contributes. Let $A_s = \sum_{i=1}^{t}\alpha_{s,i}$. The cumulative candidate exponents therefore differ by at most $2A_s$. Applying the exponent and normalization bounds from Appendix C gives

$$\frac{\Pr[S_c(T(X))=a]}{\Pr[S_c(T(X'))=a]} \le \exp(4A_s) \le \exp(\varepsilon_c).$$

Summing over output events establishes document-level privacy, and sequential composition gives the bound across categories. □

This extension uses the bounded candidate-update range and the same sampled-release calibration as the chunk-level mechanism.

G Concentration and Single-Representative Coverage

Proposition 2 (Release concentration). *Let $P = P_{s,c}^{(t+1)}$ be the final candidate distribution for a category with uniform initialization and nonnegative step sizes. Define*

$$r_c = \exp\left(2\sum_{i=1}^{t}\alpha_{s,i}\right).$$

Then $P(a)/P(a') \le r_c$ for all $a, a' \in \mathcal{A}_c$. Under Eq. (7),

$$r_c = \exp\left(\frac{\varepsilon_c}{2}\right). \tag{22}$$

With equal allocation across K nonempty categories, $r_c = \exp(\varepsilon/(2K))$.

Proof. Unrolling the synopsis updates gives

$$\log\frac{P(a)}{P(a')} = \sum_{i=1}^{t}\alpha_{s,i}\big(g_i(a) - g_i(a')\big) \le 2\sum_{i=1}^{t}\alpha_{s,i},$$

because the uniform initial weights cancel and $g_i \in [-1,1]$. Exponentiation proves the claim. □

For a fixed query routed to category c, let $N_c = |\mathcal{A}_c|$ and $n_q = \sum_{a\in\mathcal{A}_c} W(q,a)$. The single released representative satisfies

$$\Pr_{S_c\sim P}[W(q,S_c)=1] \le \frac{n_q r_c}{N_c - n_q + n_q r_c}. \tag{23}$$

To establish this bound, let p be the total probability of valid candidates. Summing $P(a) \le r_c P(b)$ over valid a and invalid b gives $(N_c - n_q)p \le n_q r_c(1-p)$. Solving for p proves the result. The cases $n_q = 0$ and $n_q = N_c$ give probabilities zero and one, respectively.

H ATTACK DESCRIPTIONS

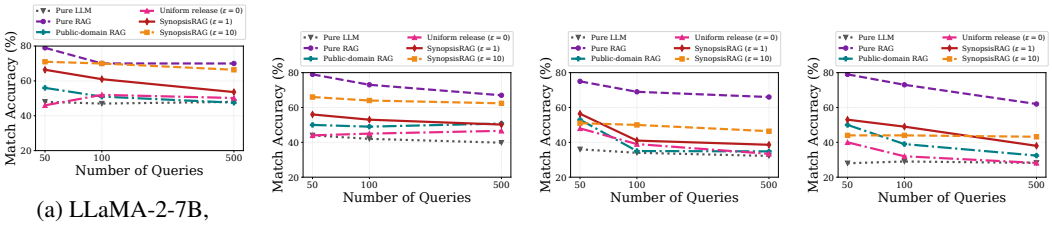
RAG-Thief (Jiang et al., 2024) is an automated, agent-based attack that extracts private RAG knowledge bases by exploiting context leakage during generation. It issues prompt-injection queries to induce partial disclosure of retrieved chunks, then iteratively learns from leaked text to generate targeted follow-up queries retrieving adjacent or semantically related content, enabling scalable reconstruction of the datastore. Pirates of the RAG (Maio et al., 2024) is a fully automated black-box attack that extracts a private knowledge base through adaptive prompt injection, iteratively deriving relevance-weighted "anchors" from leaked outputs that guide subsequent queries toward unexplored regions of the datastore. GEA (Cohen et al., 2024) escalates RAG attacks with jailbreaking to extract entire retrieved documents and propagate malicious behavior across applications, embedding adversarial self-replicating prompts that force a jailbroken model to output retrieved content verbatim and copy the payload into generated outputs. PIDE (Qi et al., 2025) exploits the retrieval-in-context design by injecting instructions that cause the LLM to copy retrieved datastore content verbatim, crafting queries that both trigger retrieval and request disclosure of hidden context, enabling scalable extraction under black-box access.

I ABLATION STUDIES

Hyperparameter ablation. Table 2 evaluates the privacy budget ε, category count K, and private retrieval size k on healthcare with LLaMA-2-7B. The reference ablation configuration uses $\varepsilon = 5$, $K = 10$, $t = 2$, $\tau = 0.7$, $\mu = 0.4$, and $k = 3$; each sweep varies one parameter while fixing the others. Each category releases one complete public representative. Accuracy increases with ε, varies modestly across K, and decreases as k increases over the tested range. Increasing k changes the private evidence used for scoring while preserving the release capacity.

Table 2: Ablation on key hyperparameters with healthcare data and LLaMA-2-7B.

Setting	Match accuracy (%)	Token leakage
$\varepsilon = 0.5$	51	136
$\varepsilon = 1$	54	139
$\varepsilon = 5$	61	144
$\varepsilon = 10$	67	155
$K = 5$	59	146
$K = 10$	61	144
$K = 15$	60	145
$K = 20$	58	147
$k = 1$	74	141
$k = 3$	61	144
$k = 5$	56	148



(a) LLaMA-2-7B, healthcare (b) Mistral-7B, healthcare (c) LLaMA-2-7B, Finance (d) Mistral-7B, Finance

Figure 5: Retrieval-source and uniform-release on healthcare and finance QA with LLaMA-2-7B and Mistral-7B.

Retrieval-source and uniform-release comparison. Figure 5 compares generation without retrieval, Pure RAG over the private corpus, public-domain RAG, uniform release, and SYNOPSISRAG on

healthcare and finance. Uniform release samples one public atom per category at $\varepsilon = 0$, using the same atom pools and generation settings as SYNOPSISRAG. Each released synopsis remains fixed throughout its evaluation sequence.

At 500 queries, Pure RAG achieves the highest match accuracy in all four dataset–model combinations. SYNOPSISRAG outperforms uniform release and public-domain RAG at both evaluated privacy budgets, with larger endpoint gains at $\varepsilon = 10$. On healthcare with LLaMA-2-7B, accuracy is approximately 54% and 67% at $\varepsilon = 1$ and $\varepsilon = 10$, respectively, compared with 50% for uniform release.

The ordering between privacy budgets varies at earlier finance checkpoints. For LLaMA-2-7B, $\varepsilon = 1$ performs better at 50 queries; for Mistral-7B, it performs better at 50 and 100 queries. Thus, the larger budget yields higher accuracy at 500 queries in both finance settings, without improving accuracy at every evaluated query count.