ROOTCERT: CERTIFIED RAG UNDER BOUNDED AUTHENTICATED-PRINCIPAL POISONING

Anonymous authors
Paper under double-blind review

ABSTRACT

A single compromised writer can corrupt retrieval-augmented generation(RAG) by introducing numerous records into external knowledge sources, making record-count budgets sensitive to fragmentation and replication. ROOTCERT addresses this issue by counting authenticated writer IDs, called *principals*, instead of records. It certifies a fixed query's reference answer against any finite append through at most r principals after an immutable memory checkpoint. Our central contribution is a *principal-to-selection reduction*. We score records separately, convexly pool them into one bounded vector per principal, and select the top-K ranked groups without merging them. If the identities, scores, and relative order of the unaffected candidates remain fixed, this deterministic procedure requires at most r matched score replacements. The standard bounded-replacement argument then ensures answer invariance when the reference answer's margin m_y over its strongest competitor exceeds $2r$. On StrategyQA, across two readers and three retrievers with frozen candidate exposure, the principal-based-certification maintains 52.6% accuracy from 1 to 64 appended records, whereas the record-based scalar certification fails after six records. A ROOTCERT-aware adaptive attacker alters 65.3% of uncertified answers at $r = 4$ without affecting any certified reference. Principal-splitting tests, additional readers, and HotpotQA are used to examine separately declared threat sets. Under these conditions, ROOTCERT safeguards certified answers against any finite record volume within a principal budget.

1 INTRODUCTION

Retrieval-augmented generation (RAG) enables pretrained language models to leverage external documents and access updated information without requiring retraining (Guu et al., 2020; Lewis et al., 2020; Izacard et al., 2023). Agents such as ReAct interleave reasoning and tool actions, while others reuse stored memories to guide later decisions (Yao et al., 2022; Shinn et al., 2023; Park et al., 2023). However, an attacker capable of adding content to these sources can influence future answers by introducing poisoned passages or memories (Zhong et al., 2023; Zou et al., 2025; Chen et al., 2024). A key question concerns the appropriate metric for the attack budget. Passage-level certificates, such as RobustRAG, constrain the number of corrupted passages (Xiang et al., 2024). However, a single compromised writer may submit multiple records. ROOTCERT addresses this discrepancy by certifying answer stability against a bounded number of authenticated writers, while permitting each writer to append any finite number of records. We represent each authenticated writer by a *principal* which is a stable ID assigned by a trusted ingestion authority. Consider a writer who submits a false claim as one passage, splits it into several chunks, or uploads repeated paraphrases. A record-level budget counts these submissions differently, whereas a principal-level budget counts one ID in every case. We fix an immutable reference memory, called the *checkpoint*, and allow the attacker to append any finite number of records through at most r principals. The attacker cannot modify the checkpoint. Our goal is a *pointwise robustness certificate*; that is, for a fixed query and checkpoint, the reference answer must remain unchanged under every append within this principal budget. This matters when a security team can justify a bound on compromised credentials but cannot bound the volume of malicious text submitted through them. The budget then tracks compromised write permissions, rather than document fragmentation or submission volume. RobustRAG's isolate-then-aggregate framework certifies responses under passage corruption, including answer invariance in finite-answer settings (Xiang et al., 2024). ROOTCERT instead budgets authenticated principals and establishes

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.

when their influence remains bounded after retrieval and selection. ReliabilityRAG and PRA-RAG certify clean-subset selection and representation deviation, respectively (Shen et al., 2026; Tan et al., 2026). Refer to Appendix A and Table 4 for a detailed comparison. Unlike the passage-level budget used by RobustRAG (Xiang et al., 2024), our principal budget allows each compromised writer to submit any finite number of records. We must therefore ensure that these records cannot affect arbitrarily many contributions to the final answer. For example, one writer's records may fill the retrieval buffer and displace evidence from other groups. Combining those records into one group after retrieval does not restore the missing evidence. Similarly, allowing new records to merge existing groups can let one writer change several group scores. Therefore, the key question is *under what retrieval, grouping, and selection conditions does control of at most r principals lead to at most r selected-score replacements?*

The theoretical contribution of RootCert is the *principal-to-selection reduction*, not a new margin certificate (Figure 1). Principal-local processing limits each writer to one group and convex pooling bounds its score independently of record count. We show that deterministic, non-merging selection of the K highest-ranked groups preserves this bound when candidate identities, scores, and relative order stay unchanged outside affected groups. Thus, at most r principals require at most r matched score replacements, giving replacement radius $\rho_r \leq r$. This is the bridge to established aggregation certification (Levine & Feizi, 2020; Wang et al., 2022b), which yields the sufficient test $m_y > 2r$ for margin m_y over the strongest competitor.

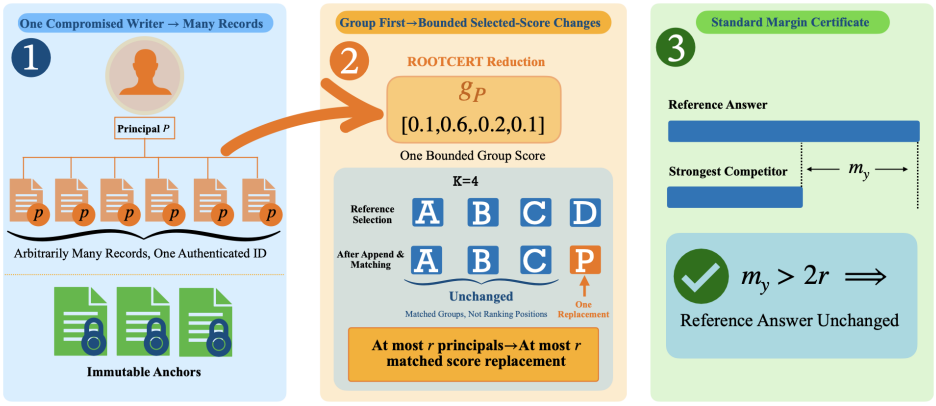

Figure 1: RootCert: Convex pooling gives each authenticated principal P one bounded score g_P. The illustrative $K = 4$ example shows P displacing D. Under the conditions in Sections 2-3, at most r principals induce at most r matched score replacements; $m_y > 2r$ preserves the reference answer.

Authentication verifies which principal submitted a record, not the truth of the content or independent ownership. Deployments must therefore justify the principal budget over the certificate's horizon. Restricted enrollment may support such a bound, whereas unrestricted enrollment permits *Sybil identities*, or multiple valid IDs controlled by one actor (Douceur, 2002). Therefore, the guarantee is limited to the declared budget and candidate-exposure scope. Our main empirical test separates record volume from principal control. One-principal certified accuracy stays at 52.6% through 64 records, whereas splitting those records across more principals weakens certification (Figure 2). ROOTCERT-aware adaptive attacks challenge uncertified answers while preserving certified references (Figure 3). Shared-calibration controls show that the accuracy gap over record accounting cannot be attributed to grouping alone (Tables 1-2). The benefit is protection against a broader set of appends, not a guaranteed accuracy gain. Our main contributions are as follows:

- We formulate principal-bounded append poisoning for RAG. The budget counts authority-assigned writer IDs rather than records, allowing any finite number of records from those IDs after an immutable checkpoint.
- We prove that appends through at most r principals require at most r matched score replacements ($\rho_r \leq r$), provided pooling is principal-local, unaffected candidates exposure stays fixed, and selection is stable and non-merging. This principal-to-selection reduction

enables the standard margin certificate $m_y > 2r$. We also identify selection rules that violate the reduction and establish the scalar margin ceiling $m_y \leq K + \lambda$.

- We evaluate ROOTCERT on StrategyQA using record replication, principal-splitting, adaptive attacks, and shared-calibration comparisons. Additional readers, HotpotQA, and tests of the underlying assumptions examine the guarantee's scope and limitations.

2 PROBLEM FORMULATION AND THREAT MODEL

Consider a RAG system answering the query q from the reference memory M, with finite answer space $\mathcal{Y}$ and abstention $\perp$. Fixing the model, prompt, retrieval, scoring, tie-breaking, decoding, and random seed makes $D(q, M) \in \mathcal{Y} \cup \{\perp\}$ deterministic. After this reference decision, the system may accept a finite append $A = (x_1, \ldots, x_n)$, giving memory $M \parallel A$ without deleting or modifying M. For fixed non-abstaining $y = D(q, M)$, we ask whether every allowed append preserves y. Each selected memory group contributes one bounded probability vector. A *trusted anchor* is a group fixed at the checkpoint. Here, "trusted" means immutable to the append attacker, not factually correct.

Let $\mathcal{R}_{\text{wr}}(M) \subseteq M$ denote reference records accepted through the writable interface, and $\mathcal{P}$ the valid-ID space. The ingestion authority assigns each accepted record x a verified ID $\texttt{pid}(x) \in \mathcal{P}$. For each ID $u \in \mathcal{P}$, the corresponding principal group G_u, whose records at memory M are

$$G_u(M) = \{x \in \mathcal{R}_{\text{wr}}(M) : \texttt{pid}(x) = u\}. \tag{1}$$

The ID comes from an authenticated session, application programming interface (API) or device credential, or signing key, not record text or unauthenticated metadata. Before global selection, each nonempty group G_p is pooled into one vector $g_p(q) \in \Delta^{\mathcal{Y}}$, the probability simplex over $\mathcal{Y}$. Thus, arbitrarily many records from one principal can change its single group score but cannot create independent scores. Authentication secures this record-to-principal mapping.

Threat Model: Principals are security units because write permissions are enforced through identities such as an employee account writing enterprise RAG documents, an API credential writing agent memories, or a service account contributing records. Fragmenting one malicious statement into many records does not increase the number of compromised IDs. Therefore, fragmentation or replication may change the pooled score or selection priority, but under the conditions stated below, it cannot multiply the principal's contribution to the selected-score replacement bound.

Let $\mathfrak{A}(M)$ be the set of finite appends accepted after M is fixed. For $A = (x_1, \ldots, x_n) \in \mathfrak{A}(M)$, define $P(A) = \{\texttt{pid}(x_i) : 1 \leq i \leq n\}$. For a fixed principal budget $r \in \mathbb{N}_0$, the allowed threat set is

$$\mathcal{A}_r(M) = \{A \in \mathfrak{A}(M) : |P(A)| \leq r\}. \tag{2}$$

Here, $\mathbb{N}_0$ denotes the nonnegative integers, and r bounds distinct verified IDs, not records. The white-box attacker knows the full pipeline and may choose each record's content, retrieval key, submission time, and unauthenticated metadata. For certification, we conservatively allow each affected principal's pooled score to range over the probability simplex. The attacker cannot modify checkpoint records or anchors, or obtain acceptance under an uncontrolled ID except with negligible cryptographic failure probability. Apart from these constraints, append construction is unrestricted. Pre-checkpoint poisoning, denial-of-service, and appends exceeding r principals are outside the threat model. Authentication binds records to IDs but does not prevent one actor from obtaining multiple IDs. Deployments must justify r over the certification horizon; otherwise, the guarantee describes only the declared threat set in Equation 2.

Example Deployment Interpretation: For example, consider a closed enterprise RAG system with centrally issued writer IDs, no self-enrollment, and daily checkpoints. Suppose that its security team can justify that attackers can write through at most two distinct IDs during the 24 hours between checkpoints. The system then uses $r = 2$ when certifying answers. The two-ID limit applies to the entire day, not separately to each attack. Revoking an ID does not remove it from the count because its earlier records remain in memory. If attackers use a third ID, their combined appends fall outside the guarantee for $r = 2$.

The system scores each group, selects the top-K groups by deterministic priority, and aggregates their vectors, with fixed $K \geq 1$. To relate the principal budget r to the selected scores, we require following

four conditions: (i) an append cannot remove or modify any record already in immutable checkpoint M; (ii) the ingestion authority assigns the verified principal ID; (iii) records from one principal can change only that principal group's summary, score, and selection priority; and (iv) candidate identity sets outside the directly affected groups are unchanged, as are their answer-score vectors and deterministic relative selection order, and an append cannot merge or split existing groups. Conditions (i)-(iii) ensure that an append changes only the groups belonging to the principals submitting it. Condition (iv) requires the other candidate groups to remain available for selection, with unchanged scores and relative order. Authentication verifies the principal IDs, but the system must enforce the remaining requirements. For example, adding records to one group must not change another group's pooling weights. In addition, keeping only the highest-ranked passages before grouping may let one principal's records displace other groups, violating (iv). These requirements govern how the system processes records; the attacker remains free to choose their content.

Let $J(A)$ be the set of group identities directly created or modified by append $A \in \mathcal{A}_r(M)$. Under conditions (ii) and (iii),

$$J(A) \subseteq \{G_u : u \in P(A)\}, \qquad |J(A)| \le |P(A)| \le r. \tag{3}$$

Section 3 transfers this bound through top-K selection.

Let $C_K(q, M) = \{c_1, \ldots, c_K\}$ be the multiset of group-score vectors selected for query q. If fewer than K groups are selected, we pad the multiset with zero vectors (empty score slots). For two padded multisets C and C', we define their replacement distance as

$$d_{\text{rep}}(C, C') = \min_{\pi \in \mathfrak{S}_K} \sum_{i=1}^{K} \mathbf{1}\left\{ c_i \neq c'_{\pi(i)} \right\}, \tag{4}$$

where $\mathfrak{S}_K$ is the set of permutations of $\{1, \ldots, K\}$ and $\mathbf{1}\{\cdot\}$ is the indicator function. This distance is the smallest number of unequal matched vectors. A group entering top-K and displacing another therefore contributes at most one replacement, not a separate insertion and deletion.

We define the post-selection radius of $\mathcal{A}_r(M)$ by

$$\rho_r = \sup_{A \in \mathcal{A}_r(M)} d_{\text{rep}}(C_K(q, M), C_K(q, M \parallel A)), \tag{5}$$

the largest change at the final scoring interface. Conditions (i)-(iv) give $\rho_r \le r$. Therefore, merely counting the groups in the final top-K is not enough to obtain this bound. An approximate retriever may hide an old candidate and violate (iv). Such exposure changes require a uniform bound $h \in \mathbb{N}_0$ on additional replacements, giving radius $r + h$ (Corollary 2). This bound must hold for every allowed append, not merely on average over tested attacks. Our experiments use $h = 0$ only within their frozen candidate scope (Section 4.1).

Certification Goal: For the fixed reference answer $y = D(q, M)$, the certificate aims to prove

$$D(q, M \parallel A) = y \quad \text{for every } A \in \mathcal{A}_r(M). \tag{6}$$

The pointwise guarantee in Equation 6 follows from $\rho_r \le r$ and the reference margin $m_y > 2r$, or $m_y > 2(r + h)$ with exposure radius h. A correct reference answer is then certifiably correct. This guarantees neither availability nor record truth, intermediate reasoning, or tool effects. The claim must specify the checkpoint, the principal budget r, the horizon, the identity authority, the exposure assumptions, and the decoding state (see Appendix D.1).

3 From Principal Control to Bounded Score Replacement

The central result is the reduction from authenticated-principal control to bounded selected-score replacement. We first define bounded group scores, then prove that the principal budget survives top-K selection. Only after establishing $\rho_r \le r$ do we apply the standard aggregation margin argument (Levine & Feizi, 2020; Wang et al., 2022b). Authentication supplies attribution; locality and stable selection supply the influence bound.

3.1 One bounded score per group

Fix a finite answer space $\mathcal{Y}$ with at least two elements. Reading record x in isolation produces a probability vector $v(x, q)$. For open answers, Appendix B.2 reduces the unbounded string space to finitely many competitors. For the anchor or principal group G_i, pool

$$g_i(q) = \sum_{x \in G_i} \alpha_{i,x}(q) v(x, q), \qquad \alpha_{i,x} \geq 0, \qquad \sum_{x \in G_i} \alpha_{i,x} = 1. \tag{7}$$

The weights may depend on the relevance or confidence calculated within G_i. Hence, g_i remains a probability vector regardless of whether the group contains one or more records. Reindex the selected scores as $g_1, \ldots, g_K$ and add them to a fixed answer distribution without retrieved records $p(q) \in \Delta^{\mathcal{Y}}$, called the trusted prior and weighted by λ:

$$F(q, M) = \lambda p(q) + \sum_{i=1}^{K} g_i(q), \qquad \lambda \geq 0, \tag{8}$$

with missing positions equal to zero. In exact arithmetic, the decision is the unique largest coordinate and abstains on a tie. For reference answer y, define

$$m_y = F_y - \max_{c \neq y} F_c. \tag{9}$$

3.2 Append-stable top-K selection

For an append A, let $J(A)$ contain the identities of groups it directly creates or modifies. An unaffected group displaced from top-K is not added to $J(A)$; its score can be matched with an incoming selected score.

Theorem 1 (**Stable Top-K Selection**). *Let $J(A)$ contain all group identities created or modified by append A. Suppose the candidate sets outside $J(A)$ coincide before and after the append, with unchanged score vectors and deterministic relative selection order. Then*

$$d_{\mathrm{rep}}\left(C_K(q, M), C_K(q, M \parallel A)\right) \leq |J(A)|.$$

For intuition, include common zero-score padding and let a, a' count affected groups selected before and after the append. The selected unaffected groups are prefixes of the same fixed order, so at least $K - \max(a, a')$ vectors match. Thus at most $\max(a, a') \leq |J(A)|$ replacements remain. This counts unmatched score vectors, not changes in rank positions.

For example, with $K = 4$, suppose that 100 records from a new principal $u \in \mathcal{P}$ pool into g_u, which displaces g_4. The selected scores change from $\{g_1, g_2, g_3, g_4\}$ to $\{g_1, g_2, g_3, g_u\}$, giving one replacement. Additional records may change g_u and its rank, but not the one-principal bound.

Mutable passage-level top-d cutoffs can evict unaffected candidates before grouping; connected-component routing can let one bridging record merge arbitrarily many groups. Appendix B.4 proves the theorem and both failure cases. Section 4.6 demonstrates the exposure failure experimentally. A uniform bound h on additional selected-score replacements gives the effective radius $r + h$.

Corollary 2 (**Exposure-Augmented Certificate**). *Let $r, h \in \mathbb{N}_0$. Suppose that for every $A \in \mathcal{A}_r(M)$ there exists a size-K score multiset $\widetilde{C}_A$ satisfying*

$$d_{\mathrm{rep}}(C_K(q, M), \widetilde{C}_A) \leq r, \qquad d_{\mathrm{rep}}(\widetilde{C}_A, C_K(q, M \parallel A)) \leq h.$$

Then $\rho_r \leq r + h$, and $m_y > 2(r + h)$ implies Equation 6.

3.3 Principal-to-selection reduction

Authentication and principal-local processing give $|J(A)| \leq |P(A)|$. Theorem 1 transfers this bound through selection, establishing the main reduction.

Proposition 3 (**Principal-to-Selection Reduction**). *Under conditions (i)-(iv) of Section 2 and correct record attribution, every $A \in \mathcal{A}_r(M)$ satisfies*

$$d_{\mathrm{rep}}(C_K(q, M), C_K(q, M \parallel A)) \leq |J(A)| \leq |P(A)| \leq r. \tag{10}$$

Hence $\rho_r \leq r$. Under the protocol and cryptographic assumptions in Appendix B.5, every probabilistic polynomial-time adversary has negligible attribution-failure probability during its execution.

The reduction establishes the replacement bound, the answer certificate then follows from Theorem 4. Signed receipts bind the assigned ID, record hash, and log position, while Merkle proofs verify checkpoint membership (Bernstein et al., 2012; Merkle, 1988). These implement attribution, not locality or selection assumptions. Authentication also does not bound how many valid IDs one actor controls; Appendix B.5 states the additional issuance assumptions needed for such an interpretation.

3.4 Applying the standard margin certificate

Given a replacement bound, we apply the established aggregation argument (Levine & Feizi, 2020; Wang et al., 2022b). Replacing one bounded score vector can reduce the reference answer's margin over a competitor by at most two. We state its score-space form for completeness.

Theorem 4 (Standard Bounded-Replacement Certificate). *For the aggregate in Equation 8, let $y = D(q, M) \in \mathcal{Y}$, $\rho \in \mathbb{N}_0$, and $\mathcal{T} \subseteq \mathfrak{A}(M)$. If $d_{\mathrm{rep}}(C_K(q, M), C_K(q, M \parallel A)) \leq \rho$ for every $A \in \mathcal{T}$, then*

$$m_y > 2\rho \quad \Longrightarrow \quad D(q, M \parallel A) = y \quad \forall A \in \mathcal{T}. \tag{11}$$

Proposition 3 supplies $\rho_r \leq r$, so setting $\rho = r$ yields the sufficient condition $m_y > 2r$. This margin test is a consequence of the reduction, not a separate new claim. Appendix B.3 gives its tightness and integer-radius calculation.

The score-aware verifier in Appendix B.2 uses individual selected vectors and is exact for arbitrary score replacements. At radius one, it adds two certified-correct primary decisions, so we retain the simpler scalar condition for the main scope comparisons.

3.5 Two limits

Corollary 5 (Structural Ceiling). *If Equation 8 contains $L \leq K$ nonzero group scores, then*

$$m_y \leq L + \lambda \leq K + \lambda. \tag{12}$$

The scalar condition $m_y > 2r$ is impossible for $r \geq \left\lceil \frac{K+\lambda}{2} \right\rceil$.

This is an algebraic boundary, not an empirical discovery. For $K = 10$ and $\lambda \in \{1, 2\}$, the scalar certificate cannot hold when the principal budget is $r \geq 6$. The second limit concerns unrestricted appends.

Theorem 6 (Unrestricted Append-Only Invariance). *Fix the query and trusted state. Let $\mathsf{M}_{\mathrm{fin}}$ be the set of finite multisets of untrusted records and $D(q, \cdot) : \mathsf{M}_{\mathrm{fin}} \to \mathcal{Y} \cup \{\perp\}$ be deterministic. If*

$$D(q, M) = D(q, M \parallel A) \qquad \forall M, A \in \mathsf{M}_{\mathrm{fin}},$$

then $D(q, \cdot)$ is constant on $\mathsf{M}_{\mathrm{fin}}$.

The premise requires invariance from every finite reference memory, so the result permits certificates at selected checkpoints. Any rule depending on untrusted memory admits an append that changes its output at some checkpoint.

4 Experiments

We compare record and principal threat sets, examine calibration sensitivity, and test replication, principal splitting, and RootCert-aware attacks. Composition and external-validity tests assess the budget and exposure assumptions. These experiments evaluate implementation and attack behavior; the guarantee against every allowed append follows from Section 3 under its stated assumptions.

4.1 Experimental setup

StrategyQA (Geva et al., 2021) uses 49 validation and 180 held-out test questions. Crossing Qwen2.5-1.5B and Mistral-7B readers with dense passage retrieval (DPR), BGE-base-en-v1.5 (BGE), and E5-base-v2 (E5) gives $1,080$ decisions per method. Each query retains 100 checkpoint candidates from AgentPoison's $9,251$-paragraph corpus. Appends leave this buffer fixed, so $h = 0$ applies only

Table 1: StrategyQA results over 1,080 held-out decisions (accuracies in %). Rows show the threat set and radius; four attack records are assigned to one or four principal IDs.

System & Threat Set	ρ	Clean	Reference	Attacked	Certified	n	Fail.
Four-record threat	4	59.1	54.6	55.9	24.3	262	0
Four principals (**RootCert**)	4	58.0	54.7	55.4	41.9	453	0
One principal (**RootCert**)	1	58.0	54.7	54.7	52.6	568	0

within that scope. Records are scored separately using at most 768 input tokens and No/Yes/Unknown next-token scores. The primary experiment pools at most four records per group, selects $K = 10$ groups, and uses validation-selected prior weight $\lambda \in \{1, 2\}$ (Appendix C.1). Checkpoint groups are immutable Wikipedia-page anchors; attack groups use synthetic authority-assigned IDs, with receipts audited separately in Appendix C.14. The primary append has four fixed AgentPoison records, evaluated at record radius $\rho = 4$ and principal budgets $r \in \{1, 4\}$.

Clean, *reference*, and *attacked* accuracy use untriggered checkpoint, triggered checkpoint, and triggered appended queries, respectively. *Certified accuracy* counts correct reference answers passing the scalar test; n is their count, and "Fail." counts changes among them under the tested append. Target attack success rate (ASR) conditions on non-target references. Overall flip rate uses all decisions; adaptive flip rates condition on uncertified references. Confidence intervals (CIs) resample questions jointly across configurations 10,000 times. We reaggregate saved scores with exact bounded-score arithmetic and abstention tolerance $\tau = 10^{-12}$, testing $m_y > 2\rho + \tau$. Unknown is an answer class, distinct from abstention $\perp$. Appendix C specifies the replay scope and numerical conventions, including the separate floating-point calibration diagnostic below.

4.2 Certified accuracy under the declared threat sets

Against four principals, 453 of 1,080 RootCert reference decisions are correct and certified (41.9%), versus 262 (24.3%) for the four-record baseline under separately selected calibration (Table 1). Within the same authenticated ingestion interface, the principal threat includes every four-record append and permits any finite number of records from those IDs. Both scalar tests use replacement radius four, and reference accuracies are similar (54.7% and 54.6%). The comparison combines grouping and calibration choices, which we examine below. At one principal, certified accuracy is 52.6%. This threat permits arbitrary finite record counts from one ID, whereas the four-record threat permits writes through several IDs; neither contains the other. Attacked accuracy describes only the evaluated append and includes uncertified answers. Appendix C.1 gives the per-configuration results; Appendix C.6 treats ReliabilityRAG separately because it certifies a different outcome.

Calibration Diagnostic: Applying either method's selected calibration to both narrows the principal-record gap from 17.69 percentage points to 1.94 with RootCert settings and 0.46 with record settings (Table 2). These archived floating-point replays retain each method's selected evidence and raw scores, with no new tuning, retrieval, or neural inference. The transferred parameters are the class biases, temperature, pooling-weight parameters, and prior weight (Appendix C.1). Shared settings need not be optimal for both systems. The diagnostic establishes calibration sensitivity, not an exact contribution from grouping. Principal accounting broadens the allowed append set without guaranteeing higher certified accuracy.

Table 2: Calibration sensitivity on the same 1,080 StrategyQA decisions. Certified accuracy (%) uses the scalar test at radius four. Gaps and paired question-bootstrap 95% intervals are in percentage points, computed before rounding the accuracies.

Calibration	RootCert	Record	Gap	95% CI
Separately selected	41.9	24.3	17.69	[14.81, 20.65]
Shared RootCert settings	41.9	40.0	1.94	[0.93, 2.96]
Shared record settings	24.7	24.3	0.46	[−0.56, 1.39]

4.3 RECORD REPLICATION AND PRINCIPAL SPLITTING

We vary record volume and principal count separately (Figure 2). An ordered 64-record Yes-target attack is optimized against passage accounting on validation data, then frozen. For an a-record prefix, passage accounting uses $\rho = \min(a, K)$; ROOTCERT pools all appended records under one ID, without the four-record cap, and uses $\rho = 1$. ROOTCERT retains the same 568 certified-correct references (52.6%) through 64 records because the reference and principal budget are fixed. Uncertified answers can still change; the replay finds zero certified violations. Passage scalar certification vanishes at six records; at eight, passage ASR reaches 68.8%, versus 0.8% for ROOTCERT. The ROOTCERT rate measures attack transfer, motivating the direct attack below. Splitting the same 64 records across one to eight principals instead raises ASR from 0.5% to 37.2% (right). Scalar certification vanishes at $r = 6$ because $m_y \leq K + \lambda \leq 12$, but an uncertified answer is not necessarily attacked successfully. ASR uses 984 passage and 991 principal non-Yes references; other rates use all 1,080 decisions. Full sweeps are in Appendix C.7.

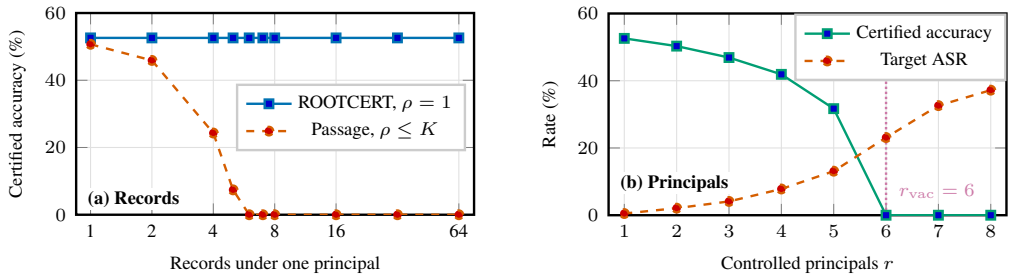

Figure 2: Replication within one principal leaves the RootCert radius unchanged (left), while splitting the same records across more principals weakens certification (right).

4.4 ROOTCERT-AWARE ADAPTIVE ATTACKS

The white-box attacker directly targets ROOTCERT, knowing the reference, calibration, pooling weights, and top-K priorities. It searches 100 payloads and 76 keys without ground-truth labels, appending at most 64 records through $r \in \{1, 2, 4\}$ new principals. At $r = 4$, the adaptive attack changes 209/320 (65.3%) certificate-failing references but 0/760 certificate-passing references, showing substantial effectiveness outside the certified set while no tested attack changes a certificate-passing output. This finite evaluation is a sanity check; the guarantee follows from Section 3, not from empirical attack failure. The transfer flip rate is 26.9% (Table 17). The exact follow-up search repeats one payload/key pair across new principals, gains level off at four keys but continues through 100 payloads (Figure 4). This finite-family search does not establish unrestricted attack optimality. Ideal pooled scores allow ten further flips; improving rank alone does not. These diagnostic relaxations are not realized text attacks (Appendix C.9).

4.5 BUDGET COMPOSITION AND UNDER-DECLARATION

The budget must cover the full certification horizon, not restart after each write. We append four frozen Unknown-target batches, each with four records under a new principal. Every prefix uses the original checkpoint; the endpoint equals the combined append. The budget counts distinct IDs, $|\bigcup_t P(A_t)|$, across appends A_t. Counting all affected IDs gives zero observed violations. At four principals, the score-aware test certifies 468 correct references (43.3%), versus 453 (41.9%) for the scalar test. Charging four principals as one instead changes four threshold-passing answers; declaring a compromised principal as $r = 0$ changes seven. These controls lie outside $\mathcal{A}_r(M)$ and show why every affected ID must be counted. Appendix C.3, Table 8, gives the complete results.

4.6 EXTERNAL VALIDITY

We test portability across readers and an open-answer task (Table 3). Llama-3.2-3B and GPT-OSS-20B tests reuse the primary split and retrievers (1,080 decisions). Qwen2.5-14B uses a 100-validation/600-test split disjoint from the primary and attack-construction sets (1,800 decisions).

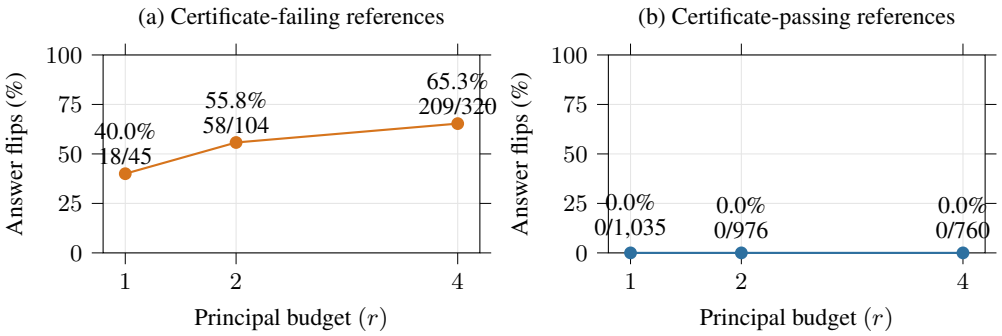

Figure 3: Adaptive attack outcomes on 1,080 decisions. Labels show flips/population; "certified" denotes certificate-passing references irrespective of correctness. At $r = 4$, 760 references pass the certificate, of which 453 are correct (41.9% certified accuracy). Non-flips among certificate-failing references are empirical attack failures, not certificates.

Table 3: Certified accuracy (%) across settings. Independent accounting counts four StrategyQA records or eight HotpotQA provider trajectories; principal accounting groups them under one ID.

Evaluation	Decisions	Independent units	Principal threat	n	Fail.
StrategyQA, Llama/GPT	1,080	21.5	43.4	469	0
StrategyQA, Qwen–14B	1,800	14.7	24.5	441	0
HotpotQA, ReAct	600	0.0	13.7	82	0

HotpotQA (Yang et al., 2018) crosses the primary readers with BGE/E5 on 30 validation and 150 test questions (600 decisions). Each provider contributes one ReAct short-answer vote from its retrieval index. Seven benign votes stay fixed; eight malicious-provider votes are averaged under one principal ($\rho = 1$) or counted independently ($\rho = 8$). Principal accounting certifies 13.7% exact-match accuracy; independent accounting certifies none. Low clean accuracy limits utility (Section 5; full results in Appendix C).

Why Retrieval Order Matters: One principal causes up to ten selected-score replacements with exact passage-first retrieval, versus at most one with exact group-first under the same tested appends (Figure 5; Appendix C.12). Principal accounting alone is insufficient. Pooling after a passage cutoff cannot recover discarded evidence. Group-first retrieval avoids this bottleneck, illustrating the exposure condition in Section 3.2. These finite observations do not certify corpus-wide retrieval.

5 Conclusion

RootCert certifies append-only RAG using authenticated principals rather than records as the corruption unit. Under explicit locality and selection assumptions, our *principal-to-selection reduction* bounds selected-score replacements by r, enabling standard margin certification against any finite record volume within that principal budget. Replication and principal-splitting experiments support this distinction. Adaptive attacks change uncertified answers without violating certified references in the tested cases. Shared calibration clarifies that broader protection does not guarantee higher certified accuracy.

Deployment requires justified principal budgets. Authentication alone establishes neither identity scarcity nor truth. Primary certificates use frozen candidate exposure, while corpus-wide retrieval requires a uniform exposure bound h. Certification guarantees stability, not correctness, and excludes pre-checkpoint poisoning and tool effects. Low clean accuracy 14.5% HotpotQA exact match and 47.9% for the 14B reader along with ownership concentration and isolated reading, limits utility. Because RootCert scores records independently before pooling, the certificate bounds integrity influence, not computational work. Arbitrarily many records from one principal may still increase latency and cost, and therefore remain an availability concern. Matched latency is unevaluated, and improving efficiency is future work.

Reproducibility statement

Appendix B provides complete proofs of all theoretical results. Section 4.1 and Appendix C specify the datasets, splits, models, retrievers, prompts, calibration procedure, threat sets, attacks, and evaluation metrics.

AI use statement

We have used LLMs to assist with writing the manuscript and experimental code debugging, and to brainstorm theoretical proof ideas. The authors retain responsibility for the methods, results, and references.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Daniel J. Bernstein, Niels Duif, Tanja Lange, Peter Schwabe, and Bo-Yin Yang. High-speed high-security signatures. *Journal of Cryptographic Engineering*, 2:77–89, 2012. doi: 10.1007/s13389-012-0027-1.

Xiaoyu Cao, Zaixi Zhang, Jinyuan Jia, and Neil Zhenqiang Gong. Flcert: Provably secure federated learning against poisoning attacks. *IEEE Transactions on Information Forensics and Security*, 17: 3691–3705, 2022.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. *Advances in Neural Information Processing Systems*, 37:130185–130213, 2024.

John R Douceur. The sybil attack. In *International workshop on peer-to-peer systems*, pp. 251–260. Springer, 2002.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-rag: certified generation risks for retrieval-augmented language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defense against general poisoning attacks. *arXiv preprint arXiv:2006.14768*, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Yedidel Louck. Securing llm-agent long-term memory against poisoning: Non-malleable, origin-bound authority with machine-checked guarantees. *arXiv preprint arXiv:2606.24322*, 2026.

Ralph C. Merkle. A digital signature based on a conventional encryption function. In *Advances in Cryptology—CRYPTO '87*, pp. 369–378. Springer, 1988. doi: 10.1007/3-540-48184-2_32.

Meta AI. Llama 3.2 model card. Meta Llama Models, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md.

Ciyan Ouyang and Rui Hou. MemLineage: Lineage-guided enforcement for LLM agent memory. *arXiv preprint arXiv:2605.14421*, 2026.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Tarun Sharma. SMSR: Certified defence against runtime memory poisoning in persistent LLM agent systems. *arXiv preprint arXiv:2606.12703*, 2026.

Zeyu Shen, Basileal Imana, Tong Wu, Chong Xiang, Prateek Mittal, and Aleksandra Korolova. Reliabilityrag: Effective and provably robust defense for rag-based web-search. *Advances in Neural Information Processing Systems*, 38:45662–45702, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Xue Tan, Yi Zheng, Chang Huo, Yunruo Zhang, Yu Liu, Hao Luan, Zhuyang Yu, Jun Dai, Xiaoyan Sun, and Ping Chen. Pra-rag: Provably robust aggregation in retrieval-augmented generation against retrieval corruption. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 36002–36017, 2026.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022a.

Wenxiao Wang, Alexander J Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (deterministic) finite aggregation. In *International Conference on Machine Learning*, pp. 22769–22783. PMLR, 2022b.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556*, 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 641–649, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13764–13775, 2023.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to Retrieval-Augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844, Seattle, WA, August 2025. USENIX Association. ISBN 978-1-939133-52-6.