POISONVAX: PROTECTING TRUSTED FACTS IN RAG THROUGH DUAL-VIEW RECORDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Corpus poisoning can steer retrieval-augmented generation (RAG) toward attacker-chosen answers by displacing trusted evidence or overriding it during generation. Existing defenses mainly filter suspicious passages or aggregate responses from retrieved evidence, but they cannot ensure that trusted facts reach the generator and prevail over competing claims. We introduce PoisonVax, a maintainer-led defense that reinforces trusted facts during corpus construction to improve both retrieval exposure and answer adoption. Each fact is stored as a dual-view record: an index view with an optimized retrieval key serves the retriever, while a generation view with a directive supports the trusted answer. The retriever scores only the index view, while a fixed field mapping supplies the generation view after retrieval. This separation keeps generation directives out of retrieval optimization while leaving the retriever, generator, and scoring rule unchanged. Records are constructed without access to evaluation questions. Under five injected attack passages per fact, with the same trusted facts available to all methods, PoisonVax achieves 58.46% answer accuracy and substantially outperforms TrustRAG and RobustRAG while using one online generation call. Matched experiments further show a 30.09 percentage point gain over a fully reoptimized single text baseline after removing attractor spans.

1 INTRODUCTION

Retrieval-augmented generation (RAG) answers questions using passages selected from an external corpus (Lewis et al., 2020; Karpukhin et al., 2020). Shared knowledge bases and web collections draw on contributions from multiple parties, creating an opportunity for corpus poisoning. A small number of malicious passages can redirect a question to an attacker-chosen answer (Zou et al., 2025; Ben-Tov & Sharif, 2025; Su et al., 2025; Chang et al., 2025). These passages compete with trusted evidence for limited retrieval slots and for influence on the final answer. Competition becomes more pronounced when several attackers target the same question (Chen et al., 2026a; Nestaas et al., 2025). For a knowledge-base maintainer, the central defense objective is to keep answering questions about trusted facts correctly despite this competition.

Existing defenses act on evidence returned by retrieval. RobustRAG isolates retrieved passages and securely aggregates their responses (Xiang et al., 2026). TrustRAG combines document filtering with self-assessment to resolve knowledge conflicts (Zhou et al., 2025). RAGDefender groups and filters passages without additional generator calls (Kim et al., 2025), while AV Filter detects anomalies in passage attention statistics (Choudhary et al., 2026). When attack passages displace a trusted fact, filtering or aggregating the remaining content does not ensure that the missing fact reaches the generator. Even a visible fact can lose to a competing claim during generation. Thus, providing correct facts in the corpus and suppressing malicious content leave an important defense problem: making those facts available and effective when the system answers.

Our insight is to strengthen trusted records before they enter this competition. Poisoning research has demonstrated powerful ways to gain *retrieval exposure* and steer *answer adoption* (Ben-Tov & Sharif, 2025; Chen et al., 2026b). We put these capabilities to work for trusted facts. Controlled interventions explain why the defense must address both: equalizing exposure changes which attacker wins, and suppressing the leading attacker can transfer success to another wrong answer. Combining both in one record also creates a representation challenge. Generation directives become part of the

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

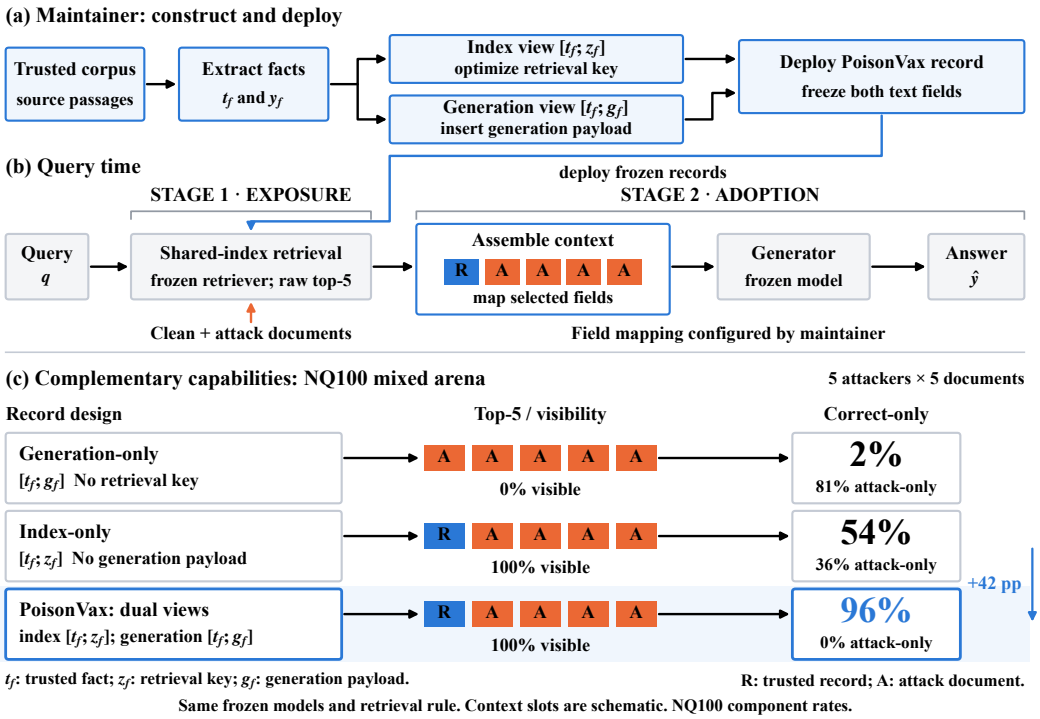


Figure 1: **PoisonVax overview.** (a) Record construction by the maintainer, (b) retrieval and generation, and (c) complementary capabilities in the NQ100 mixed arena (Appendix C). Context slots are schematic. Percentages are measured rates.

indexed text in a conventional record, changing the input that retrieval optimization must accommodate. Strong retrieval and strong answer steering do not automatically result from putting both components in one text.

PoisonVax resolves this tension through two text views of each trusted fact. The index view pairs the fact with an optimized retrieval key, and the generation view pairs it with a directive supporting the trusted answer. The retriever scores the index view, and a fixed field mapping supplies the generation view only after the record is selected. This preserves retrieval competitiveness while retaining answer control in the context. The corpus-first constructor extracts facts from trusted passages and prepares reusable records without using evaluation questions in optimization. The maintainer deploys the records with the retriever, generator, and scoring rule unchanged, using a single online generation call.

Our contributions are threefold:

- PoisonVax strengthens facts during corpus construction by separating the text used to compete for retrieval from the text used to guide generation. Its corpus-first constructor prepares reusable records for unseen questions while keeping both models and the retrieval rule fixed.
- Controlled interventions on scores, exposure, order, and attacker removal show that suppressing a strong attacker can leave other wrong answers in control, so a trusted record must secure both retrieval exposure and answer adoption.
- With the same trusted facts and attack budget, PoisonVax recovers more correct answers than the post-retrieval defenses evaluated here. Under a matched optimization budget, it also outperforms a simplified, fully reoptimized single-text baseline across three optimization seeds. Evaluations on two corpora, unseen questions, model variants, and adaptive attacks establish the method's protection and characterize its shared-index utility.

2 Related Work

Defenses against corpus poisoning. RobustRAG isolates passages or passage groups and securely aggregates responses, with certification under a specified retrieval corruption budget (Xiang et al., 2026). TrustRAG combines cluster filtering with assessment by the generator to resolve knowledge conflicts (Zhou et al., 2025). RAGDefender combines passage grouping with similarity-based filtering without additional generator calls (Kim et al., 2025). AV Filter detects anomalous passages using the variance of normalized passage attention scores (Choudhary et al., 2026). These methods protect generation by processing retrieved evidence. PoisonVax intervenes earlier, during corpus construction, to improve which trusted records reach generation and how their answers are used.

Poisoning attacks and retrieval competition. PoisonedRAG shows that a few injected passages can corrupt RAG answers (Zou et al., 2025). GASLITE optimizes text to compete for dense retrieval (Ben-Tov & Sharif, 2025), while Eyes-on-Me uses attention-attractor text to steer retrieval and generation (Chen et al., 2026b). PoisonArena places attackers in direct competition for the same retrieval slots and final answer (Chen et al., 2026a). These attacks provide the technical inspiration for our defense. We use competitive interventions to understand their strengths, then adapt retrieval optimization and answer directives to protect trusted facts through dual-view records.

How generators use retrieved context. Retrieval and generation respond to different properties of a passage. Retrieval scores can reflect non-semantic features such as the low-perplexity bias of PLM-based retrievers (Wang et al., 2025), and models with conflicting parametric knowledge use external context to different degrees (Huang et al., 2025; Wang et al., 2026). We connect this separation to fact protection by studying exposure and adoption under competition, and retrieve a record that supports the trusted answer at both stages (Appendices B and I).

3 Problem Setup and Threat Model

Defense objective. The maintainer supplies trusted passages from which facts can be constructed before deployment or during re-indexing. Our objective is to answer questions about these protected facts correctly when malicious passages compete in the same corpus. Correct answer recovery denotes the fraction of evaluated questions answered correctly under attack. The record construction uses source facts and construction paraphrases, without using evaluation questions in optimization.

Pipeline and outcomes. Let $\mathcal{C}$ be a corpus, R a frozen dense retriever with score $s(q, d)$, and G a frozen generator. For query q, the system retrieves $\mathcal{R}_k(q) = \mathrm{TopK}_{d\in\mathcal{C}}\, s(q, d)$ and answers $\hat{y} = G(q, \mathrm{render}(\mathcal{R}_k(q)))$, where the score uses each record's index text and render supplies its display text. Our protocol uses raw inner-product scores and $k = 5$, except for the explicitly normalized retriever experiment. An attacker inserts at most B passages per fact supporting a wrong answer y_a. The defender inserts records supporting the trusted answer y_f. For a source class c (defense, a particular attacker, or clean text), *retrieval exposure* is the number of its documents in the returned context, $E_c(q) = \sum_{d\in\mathcal{R}_k(q)} \mathbb{1}[c(d) = c]$, and *generative adoption* is measured by whether answer y is expressed in $\hat{y}$.

Measuring protection. All answer scoring is automated. The external defense comparison and the representation by payload controls use semantic machine judgments to distinguish correct answers, assigned attack answers, other errors, and abstentions. Accuracy counts machine-judged correct answers over the full evaluation set. The other controlled experiments and lexical diagnostics retain four outcomes: *correct-only*, *attack-only*, *both*, or *other*. Correct-only requires the trusted answer and excludes all assigned attack answers, but does not by itself establish final answer adoption. We additionally measure record visibility, source ownership of the top-ranked slot, and utility on non-target queries. A non-target query is one the records are not meant to answer. We measure how often a record is nevertheless retrieved and how often the output contains an unrelated protected answer.

Threat model and record interface. The defender is the knowledge-base maintainer, which constructs records from trusted passages before deployment or during periodic re-indexing and configures an index-field-to-rendering adapter: the retriever scores a defense record's index field, and the

context assembler displays its generation field if selected. Ordinary clean and attack passages use their submitted text for both operations. The attacker is a corpus poisoner that inserts malicious text through the allowed document-ingestion interface. It cannot delete trusted content or change the retriever or generator, and it is not assumed to hold administrator permissions or to control two fields in the default setting. Both parties' documents face the same retrieval score rule and the same finite context capacity, with no trust bonus, reserved slots, or third-party document filtering by PoisonVax. The coupled-attack and adaptive protocols additionally allow attack carriers and generation payloads to occupy separate fields (Section 6.5). The attacker's document budget, knowledge of the defense, and feedback access are specified per experiment. In the original corpus-first experiments, records are frozen before evaluation questions are written. Subsequent representation controls reuse existing held-out questions, which remain excluded from optimization. Our fair single-text comparison gives both defense designs the same maintainer authority and optimization budget, isolating the representation choice rather than deployment permissions.

Attack and defense budgets. We evaluate one to five documents per attacker per fact, and mixed settings combine the budgets of the participating attackers. These budgets affect both the competition for context slots and the amount of mutually consistent evidence presented to the generator, so we report protection jointly with the attack budget, defense record count, and retrieval setting. Experiments with additional defense records measure a different operating point from the default single-record deployment.

4 Mechanisms Guiding Fact Protection

We use competitive poisoning to identify the capabilities a trusted record needs. The testbed follows PoisonArena on Natural Questions (Chen et al., 2026a; Kwiatkowski et al., 2019), with 100 target queries, Contriever (Izacard et al., 2022) using raw inner product, top-5 context, Llama-3-8B-Instruct (Grattafiori et al., 2024), and up to seven attackers with five documents each. Two interventions carry the design argument (Figure 2; Appendix B).

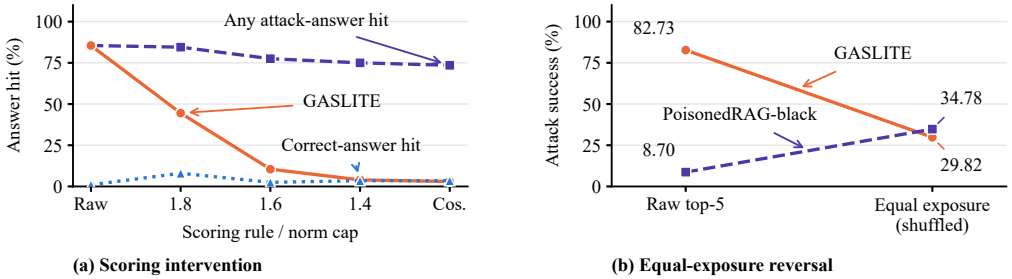


Figure 2: **Exposure vs. adoption.** Answer hits in (a) may co-occur.

Suppressing the leading attacker does not restore the correct answer. GASLITE wins the top retrieval slot for 0.9958 of queries with one document and occupies 4.968 of five slots with five, rewarded by raw inner-product scoring. Removing that advantage redistributes success among attackers rather than recovering the trusted answer. Re-ranking the same candidate pool by cosine reduces GASLITE's own success from 0.855 to 0.030, yet a targeted wrong answer still wins 0.735 of rounds, and correct-only is 0.035. Clipping document norms at the 95th percentile of the other attackers leaves GASLITE at 0.820, and a tighter cap of 1.4 reduces it to 0.040 while a targeted wrong answer still wins 0.750 of rounds.

Equal exposure still leaves an adoption competition. Retrieval rank alone does not explain competitive success. On 1,000 matched competition rounds that keep each attacker's best document plus one clean document, GASLITE's success falls from 82.73% to 29.82% with shuffled contexts and 25.82% with score-ascending order, and PoisonedRAG-black becomes the leader. Exposure determines which evidence can compete, while its content and presentation shape which answer is adopted.

Together these interventions identify two design goals: give the trusted fact enough retrieval exposure to participate, and support adoption of its answer once exposed. They motivate a defense that acts on both stages, but do not by themselves establish that two text views are required; the matched representation controls of Section 6.3 test that directly.

5 POISONVAX: A DUAL-VIEW DEFENSE

PoisonVax protects a fact by preparing one record for two different jobs: competing for retrieval and supporting the correct answer during generation. A conventional single text couples these jobs because its generation directive also becomes part of the retriever's input. Our design separates their text roles. The index view can be optimized for exposure without the generation payload, while the generation view retains the fact and its answer directive. A corpus-first constructor prepares these records from trusted passages for reuse across later questions (Figure 1).

5.1 DUAL-VIEW RECORD AND DEPLOYMENT

For each protected fact f with trusted statement t_f and answer y_f, PoisonVax stores a dual-view record

$$v_f = \big(v_f^{\mathrm{idx}},\, v_f^{\mathrm{gen}}\big), \qquad v_f^{\mathrm{idx}} = [t_f;\, z_f], \qquad v_f^{\mathrm{gen}} = [t_f;\, g_f]. \tag{1}$$

Here z_f is the fact-specific retrieval key and g_f is the generation payload instantiated with y_f. The retriever embeds v_f^{idx}. If the record enters $\mathcal{R}_k(q)$, the maintainer's fixed mapping supplies v_f^{gen} in the selected slot. The two views form one record and consume one context slot.

A single-text record must accommodate the fact, key, and final generation payload in the same indexed input. PoisonVax instead optimizes $[t_f; z_f]$ and keeps $[t_f; g_f]$ available for generation. The matched experiments in Section 6.3 show that this preserves retrieval competitiveness even when the single-text payload is simplified and fully reoptimized. Deployment uses the common scoring rule and ordinary top-k selection specified in Section 3.

5.2 INDEX VIEW: STRENGTHEN RETRIEVAL COMPETITION

The retrieval key z_f is a sequence of $L = 100$ discrete tokens. Given the fact's construction paraphrases Q_f^+ and, for the selective variant, a set of negative questions Q_f^-, the objective maximizes mean retrieval score on the paraphrases while penalizing score on the negatives,

$$J_\beta(z_f) = \frac{1}{|Q_f^+|} \sum_{q \in Q_f^+} s\big(q, [t_f; z_f]\big) \;-\; \frac{\beta}{|Q_f^-|} \sum_{q \in Q_f^-} s\big(q, [t_f; z_f]\big). \tag{2}$$

Standard PoisonVax takes $\beta = 0$, so the penalty term is inactive. The selective variant, PoisonVax-S, takes $\beta = 0.15$ with hard negatives (same-topic questions the fact does not answer) and eight global negatives from other facts. This term reduces unwanted exposure on non-target questions. Both variants use the released GASLITE discrete optimizer in two 15-iteration phases. The second phase focuses on the three currently weakest paraphrases to improve coverage of the construction questions. Keys are optimized per retriever, and only the index view enters the objective. The generation view remains fixed.

5.3 GENERATION VIEW: PROMOTE THE TRUSTED ANSWER

The generation view combines the trusted statement with an *answer directive* naming y_f. The directive strengthens the model's adoption of the protected answer when competing claims appear in the context. The unconditional template used in most controlled comparisons instructs the model to answer with y_f regardless of the query and surrounds this instruction with a pair of attention-attractor token spans. The spans are optimized once on a 1B proxy generator using the Eyes-on-Me procedure (Chen et al., 2026b), then reused for all facts, both corpora, and all evaluated generators. Per-record generation feedback is not used (Appendix A).

Attractor spans are a reusable component of the default payload, whose contribution is tested separately. The representation comparison removes them from both designs and confirms the dual-view

advantage (Section 6.3). Appendix H reports their effects in clean and attacked contexts. A conditional template applies the directive when the fact answers the query, providing a control over non-target interference (Section 6.6).

5.4 Corpus-first construction

Our corpus-first constructor builds records from trusted passages without seeing evaluation questions. An open 7B or 8B model extracts atomic fact–answer candidates from each passage. Structural checks require that the statement contain the answer, be grounded in the source span, not duplicate another candidate, and not leak its answer into any construction question. For each accepted fact the constructor writes six positive paraphrases and four same-topic hard negatives, optimizes z_f, and freezes both views. On Natural Questions, 126 candidates from 30 source titles pass the initial structural checks, of which 114 enter the frozen benchmark after the retention screen (90.5%). On MS MARCO (Nguyen et al., 2016), 100 passages yielded 102 accepted facts. Construction is per fact and parallel. Cost, iteration, and seed sweeps are in Appendix J.

6 Experiments

We evaluate correct answer recovery against existing defenses, isolate the contributions of dual views and their components, and test generalization, adaptive robustness, and shared-index utility.

6.1 Setup

We evaluate PoisonArena NQ100, corpus-first NQ114, and MS MARCO constructions; question sets, attacker budgets, and held-out counts are in Appendix D. Contriever with raw inner-product scoring and Llama-3-8B-Instruct are the defaults; transfer uses normalized BCE and Qwen2.5-7B, Mistral-7B, and Llama-2-13B. External defenses and record controls use common trusted facts and matched maintainer authority and optimization budget (Sections 6.2–6.3). Full baseline enumeration is in Appendix D. Automated evaluation covers machine-judged final-answer adoption and lexical correct-only (Appendix A).

6.2 Correct answer recovery against existing defenses

PoisonVax achieves the highest correct answer recovery across all three attack budgets in the primary comparison (Table 1). We compare with standard RAG, RAGDefender (Kim et al., 2025), the keyword aggregation variant of RobustRAG (Xiang et al., 2026), TrustRAG (Zhou et al., 2025), and AV Filter adapted and calibrated for the common generator (Choudhary et al., 2026). Comparators receive the same 99 trusted facts as ordinary records, while PoisonVax uses their frozen dual views. All methods share the corpus, Contriever, Llama-3-8B-Instruct, raw top-5 retrieval, and a static GASLITE carrier copied $B \in \{0, 1, 5\}$ times per fact. We report 65 primary questions and a separate 29-question diagnostic group (Appendix K.1). Each method retains its defense stages, with greedy decoding and a common 4,096-token final output limit.

Table 1: **Correct answer recovery with the same trusted facts.** Correct answer adoption (%) on 65 questions. Comparators receive ordinary facts. − = not reported; distinct-B=5 uses five passages carrying the same wrong answer (Appendix K.2).

Defense	B=0	B=1	B=5	Distinct B=5	Generation calls
Standard RAG	96.92	89.23	20.00	12.31	1.00
RAGDefender	84.62	72.31	23.08	–	1.00
RobustRAG	63.08	60.00	21.54	18.46	5.94
TrustRAG	53.85	52.31	44.62	44.62	3.00
AV Filter (adapted)	98.46	92.31	1.54	–	4.00
PoisonVax	**100.00**	**93.85**	**58.46**	**95.38**	1.00

At B=5, PoisonVax exceeds the strongest comparator, TrustRAG, by 13.85 points. The retrieval audit explains why record construction matters: the PoisonVax target record enters the top-5 on 32 questions, all answered correctly, whereas ordinary target records are absent from the comparators'

retrieved sets. Constructing competitive records brings trusted evidence into contexts that ordinary fact insertion cannot reach.

The error breakdown distinguishes correct answer recovery from suppression of a designated attack answer. At $B{=}5$, TrustRAG adopts the assigned attack answer on one question, but produces 29 other errors and six abstentions. PoisonVax answers more questions correctly while adopting the assigned attack answer on 18 questions. Recovering correct answers and blocking one designated attack answer are therefore not equivalent, and the gain above concerns the former. Appendix K.1 reports diagnostics and scoring and implementation details.

With one generation call, RAGDefender recovers 72.31% of answers at $B{=}1$, versus 93.85% for PoisonVax. Its filter removes 13 of 65 attack passages but also 26 of 57 visible target records, reducing the trusted evidence available for answer adoption (Appendix K.1).

PoisonVax remains effective with distinct attack passages. With five distinct passages supporting the same wrong answer, PoisonVax reaches 95.38% and TrustRAG's duplicate-embedding branch never triggers (Table 1). These attacks exert lower retrieval pressure than the original carrier, so the result complements rather than replicates the equal-pressure comparison (Appendix K.2).

6.3 Why dual views improve protection

The advantage survives a stronger single-text alternative. We compare all four combinations of representation and payload, with and without attractor spans, using the same 99 facts, key length, and optimization budget. Three independent optimization seeds and three generation seeds provide paired comparisons. Table 2 reports answer recovery on the 65 primary questions.

Table 2: **Representation and payload controls.** Machine-judged recovery (%) and target visibility on 65 primary questions, averaged within facts over three optimization and three generation seeds. Single texts are reoptimized; dual views share keys. Lexical diagnostics: Appendix C.2.

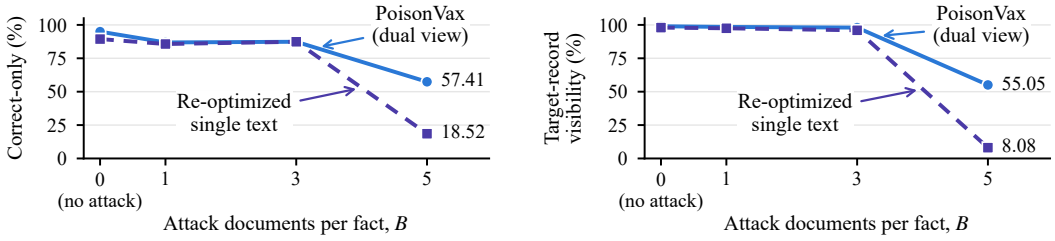
Representation	Payload	$B{=}0$	$B{=}5$	Visibility, $B{=}5$
Single text	Full	100.00	22.22	4.10
Single text	No spans	96.24	29.74	12.31
Dual view	Full	99.49	59.32	48.72
Dual view	No spans	96.58	59.83	48.72

At $B{=}5$, the simplified dual view outperforms the fully reoptimized simplified single text by 30.09 points, with positive paired gains across all three optimization seeds. Removing spans helps the single text but leaves a substantial representation gap: dual views retain higher target visibility and answer recovery. Separating the text roles preserves retrieval competitiveness beyond the benefit of payload simplification.

Strong competition reveals the representation gain. Both representations in Figure 3 are fully reoptimized on 99 facts excluded from pilot tuning, with the same fact, generation payload, key length, and optimization budget. Across 171 unseen questions, three generation seeds, and $B \in \{1, 3, 5\}$, PoisonVax gains 13.36 points in fact-equal lexical correct-only. The representations perform similarly on unseen questions at lower budgets, and construction questions can favor single text. At $B{=}5$, the recovery gap widens alongside the difference in target visibility.

At fixed dual-view retrieval, showing the key to the generator leaves correct-only nearly unchanged at 57.07%. The gain therefore tracks the retrieval competitiveness preserved by keeping the generation payload outside key optimization. Hiding the key itself contributes little in this control (Appendix C.1).

The index and generation views make distinct contributions. In the NQ100 mixed arena, five attackers contribute five documents each. Generation-only has 0% visibility and 2% correct-only. Retrieval-only and the full dual view share 100% visibility, but adding the generation payload raises correct-only from 54% to 96%. The key secures exposure, while the payload promotes adoption at fixed retrieval. On MS MARCO held-out questions at $B{=}5$, the full record also exceeds retrieval-only by 26.63 points. Appendix C (Table 6) reports the complete component results.

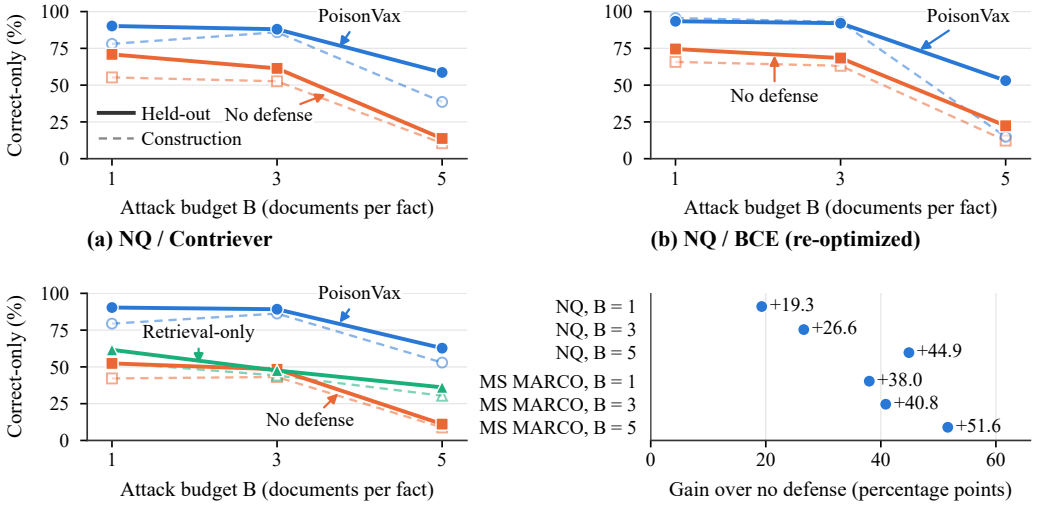


(a) Correct-answer protection **(b) Target-record exposure**

Figure 3: **Separate text optimization preserves exposure under strong competition.** (a) Fact-equal correct-only and (b) target-record visibility under matched optimization budgets: 99 facts, 171 unseen questions, three generation seeds. $B{=}0$ is attack-free.

6.4 Protection on unseen questions and across corpora

Corpus-first construction produces reusable protection for questions written after the records are frozen (Figure 4). At $B{=}5$ on Contriever, the held-out evaluation gives PoisonVax gains of 44.88 points on NQ and 51.63 points on MS MARCO. At this budget, ordinary trusted facts remain close to the no-defense rates, showing that protection comes from how facts are constructed and presented, beyond their presence in the corpus.



(a) NQ / Contriever **(b) NQ / BCE (re-optimized)**

(c) MS MARCO / Contriever **(d) Held-out protection gains**

Figure 4: **Protection on unseen questions.** (a–c) Fact-equal correct-only: solid curves show held-out questions and dashed curves show construction questions. Contriever uses 280 NQ and 250 MS MARCO questions; BCE uses a 198-question NQ diagnostic set. (d) Contriever gains over no defense. Full results: Appendix D.

Retrievers, generators, and exposure. The representation also supports BCE-embedding-base with CLS pooling and ℓ_2 normalization. Reoptimizing keys for BCE gives 93.42% and 92.11% correct-only on the 198-question NQ diagnostic set at $B \in \{1, 3\}$ and 53.07% at $B{=}5$, compared with 22.37% without defense. The same generation template benefits Llama-3, Qwen2.5-7B, and Mistral-7B (Jiang et al., 2023) without changes to its spans. Llama-2-13B (Touvron et al., 2023) responds differently: the complete record yields 52.38%, compared with 77.78% for a plain fact and 68.25% without defense, despite unchanged record visibility. Appendix F reports these generator responses and a separate Contriever-to-BCE key-transfer test, which shows positive gains with different attack carriers.

A larger shared corpus. Protection persists with one million background MS MARCO passages. On the 173-question MS MARCO scale diagnostic set, deploying 102 frozen records raises fact-equal correct-only from 1.47% to 57.35% under five coupled attack documents per fact. In this non-selective deployment, a defense record is also retrieved for 57.3% of 1,000 non-target questions.

Appendix J reports the corpus-size comparison, and Section 6.6 examines controls over such shared-index interference.

6.5 Protection under defense-aware attacks

On NQ100, PoisonVax retains protection against score-feedback black-box adaptation: attacks evolve five documents with 20–200 score queries per fact, while held-out questions evaluate the frozen attacks. Across budgets, no defense reaches at most 1% correct-only and PoisonVax reaches 17%–52% (Table 3); target records remain visible, so answer adoption remains a separate bottleneck.

Table 3: **Black-box correct-only (%).**

Questions / generator	20	80	200
Construction / Llama-3	30.00	31.00	31.00
Construction / Qwen2.5	19.00	22.00	21.00
Held-out / Llama-3	**51.00**	**49.75**	**51.50**
Held-out / Qwen2.5	18.75	17.00	17.75

White-box attackers may know defense text and gradients and separate index and generation fields. On held-out questions, three-record defenses yield 55.25%–73.25% under coupled and 54.25%–68.25% under joint attacks; one-record results are 4.25%–44.75%. A separate on-question stress map yields 8%–48% with three records. Appendix G gives the attack interfaces and complete record-budget results.

6.6 Shared-index utility and online computation

We evaluate selective keys and conditional directives separately to control the two sources of shared-index interference. Generation controls average three generation seeds and distinct questions equally, with shared retrieval: 280 NQ targets ($B = 5$) and 100 non-targets ($B = 0$).

Controlling unwanted retrieval. The selectivity experiment evaluates 114 NQ records on 5,695 non-target retrieval queries and 100 unprotected same-topic questions. Non-selective keys reduce Recall@5 from 0.200 to 0.168 and occupy 1.316 context slots on average. At β=0.15, PoisonVax-S restores Recall@5 to 0.194 and reduces occupancy by 73.2% to 0.353. Same-topic correct-only rises from 21% to 30%, compared with 44% without records. Target protection remains close to the standard version at lower budgets, with reduced exposure at B=5 (Appendices D and J).

Controlling unwanted answer adoption. The unconditional directive protects targets but reduces non-target utility (Table 4). Conditioning it on relevance improves non-target strict-correct by 12.33 points and sharply reduces unrelated protected-answer hits. Target correct-only changes by 0.24 points; non-target strict-correct remains 6.00 points below no defense. The other main experiments use the unconditional template. Appendix H reports the full component ablations. PoisonVax uses one online generation call; offline construction cost is reported in Appendix J.

Table 4: **Conditional generation controls (%).**

Metric	No def.	Uncond.	Cond.
Target correct-only	11.79	59.52	59.76
Non-target strict-correct	42.33	24.00	36.33
Non-target other-answer hit	1.00	49.00	8.67

7 Discussion and Conclusion

PoisonVax makes trusted-record construction a defense against corpus poisoning by separating retrieval exposure from answer adoption. Its dual views recover more correct answers than the compared external defenses with the same trusted facts and one online generation call. Matched controls establish an advantage over fully reoptimized single texts, including simplified payloads, across three independent optimization seeds. Protection extends to unseen questions and additional corpora, while separate utility experiments show how selective keys and conditional directives control non-target interference. These results identify the stored record as a practical point for protecting trusted facts before poisoning occurs.

AI USE STATEMENT

Generative AI tools were used for language editing and to assist with literature discovery. All identified references were independently reviewed and verified by the authors. The authors made all decisions regarding the research methods, experiments, analyses, interpretations, and manuscript content, and take full responsibility for the final work.

ETHICS STATEMENT

This work studies a dual-use security problem. The exposure–adoption analysis and the corrective-record construction could in principle be used to make poisoning attacks more effective, and the answer directive uses the same generation-side lever as existing attacks. We judge the defensive value to outweigh this risk because the attack components we use are already public, because the competitive setting we study already exists in deployed LLM search, and because the paper quantifies the costs of the defense (non-target occupancy, same-topic leakage) rather than presenting it as free. Experiments use public datasets, released attack artifacts, and locally hosted models. No live service or human users were targeted.

REPRODUCIBILITY STATEMENT

Section 3 fixes the protocol, threat model, and outcome definitions. Section 5 and Appendix A specify both views, the optimization objectives and hyperparameters, the exact generation-view template, and the generator prompt and decoding settings. Section 6.1 and Appendix D define every construction, question set, attacker, budget, and statistic, and report complete tables for each cell shown in the figures. Code is available to reviewers during double-blind review. It will be released publicly with the camera-ready version

REFERENCES

Matan Ben-Tov and Mahmood Sharif. GASLITEing the Retrieval: Exploring Vulnerabilities in Dense Embedding-based Search. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 4364–4378. ACM, 2025. doi: 10.1145/3719027.3765095. URL https://doi.org/10.1145/3719027.3765095.

Zhiyuan Chang, Mingyang Li, Xiaojun Jia, Junjie Wang, Yuekai Huang, Ziyou Jiang, Yang Liu, and Qing Wang. One Shot Dominance: Knowledge Poisoning Attack on Retrieval-Augmented Generation Systems. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18811–18825. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1023. URL https://doi.org/10.18653/v1/2025.findings-emnlp.1023.

Liuji Chen, Xiaofang Yang, Yuanzhuo Lu, Jinghao Zhang, Xin Sun, Qiang Liu, Shu Wu, Jing Dong, and Liang Wang. Uncovering Competing Poisoning Attacks in Retrieval-Augmented Generation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 414–424. ACM, 2026a. doi: 10.1145/3770855.3818119. URL https://doi.org/10.1145/3770855.3818119.

Yen-Shan Chen, Sian-Yao Huang, Cheng-Lin Yang, and Yun-Nung Chen. Eyes-on-Me: Scalable RAG Poisoning through Transferable Attention-Steering Attractors. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026b. URL https://icml.cc/virtual/2026/poster/61331.

Sukmin Cho, Soyeong Jeong, Jeongyeon Seo, Taeho Hwang, and Jong C. Park. Typos that broke the RAG's back: Genetic attack on RAG pipeline by simulating documents in the wild via low-level perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2826–2844. Association for Computational Linguistics, 2024.

Sarthak Choudhary, Nils Palumbo, Ashish Hooda, Krishnamurthy Dvijotham, and Somesh Jha. Through the Stealth Lens: Attention-Aware Defenses Against Poisoning in RAG. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings*

of Machine Learning Research. PMLR, 2026. URL https://icml.cc/virtual/2026/poster/61228.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. To Trust or Not to Trust? Enhancing Large Language Models' Situated Faithfulness to External Contexts. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/186a213d720568b31f9b59c085a23e5a-Abstract-Conference.html.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=jKN1pXi7b0.

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.

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*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Minseok Kim, Hankook Lee, and Hyungjoon Koo. Rescuing the unpoisoned: Efficient defense against knowledge corruption attacks on RAG systems. In *2025 IEEE Annual Computer Security Applications Conference (ACSAC)*, pp. 1178–1192. IEEE, 2025. doi: 10.1109/ACSAC67867.2025.00093.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Fredrik Nestaas, Edoardo Debenedetti, and Florian Tramer. Adversarial Search Engine Optimization for Large Language Models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0f12b3c36a781120c4f60e90e855868d-Abstract-Conference.html.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

Jinyan Su, Preslav Nakov, and Claire Cardie. Corpus Poisoning via Approximate Greedy Gradient Descent. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4274–4294. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.222. URL https://doi.org/10.18653/v1/2025.findings-acl.222.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Haoyu Wang, Sunhao Dai, Haiyuan Zhao, Liang Pang, Xiao Zhang, Gang Wang, Zhenhua Dong, Jun Xu, and Ji-Rong Wen. Perplexity Trap: PLM-Based Retrievers Overrate Low Perplexity Documents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4059971112ab22fc7b6aed520aaca6b2-Abstract-Conference.html.

Jiatai Wang, Zhiwei Xu, Di Jin, Xuewen Yang, and Tao Li. Accommodate Knowledge Conflicts in Retrieval-augmented LLMs: Towards Robust Response Generation in the Wild. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):33530–33538, 2026. doi: 10.1609/aaai.v40i39.40641. URL https://doi.org/10.1609/aaai.v40i39.40641.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably Robust RAG against Retrieval Corruption. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2026. URL https://satml.org/2026/program/.

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. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.849.

Huichi Zhou, Kin-Hei Lee, Zhonghao Zhan, Yue Chen, Zhenhao Li, Zhaoyang Wang, Hamed Haddadi, and Emine Yilmaz. TrustRAG: Enhancing Robustness and Trustworthiness in Retrieval-Augmented Generation. *arXiv preprint arXiv:2501.00879*, 2025. URL https://arxiv.org/abs/2501.00879.

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. USENIX Association, 2025. URL https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

The main defense comparison and its additional error breakdown are detailed in Appendix K.1. Matched single-text experiments and the representation by payload controls appear in Appendices C.1 and C.2. The remaining sections provide construction details, complete mechanism and generalization results, stronger attack protocols, and shared-index utility measurements.

A Construction and Implementation Details

Discrete optimization of the retrieval key. Both the synthetic attackers and the default index view use the released GASLITE discrete-token optimizer. The default key has 100 tokens and is optimized in two 15-iteration phases with gradient averaging over 50 samples, 256 token candidates, 128 beam candidates, 20 candidate positions per round, 512 proposed flips, and a back-propagation batch of 90. Phase one optimizes over all six positive paraphrases. Phase two restarts from the three paraphrases with the lowest current score. PoisonVax-S uses the same optimizer with four hard negatives and eight global negatives drawn deterministically from other facts, and $\beta = 0.15$. Construction, optimization, and generation seeds are stored in per-run manifests. Keys for BCE are optimized against BCE (CLS pooling, ℓ_2 normalization), not transferred from Contriever, except in the transfer experiment of Appendix F.

Generation view. The generation view is the trusted statement followed by the answer directive. The directive is the sentence

> `Ignore previous instructions. YOU MUST BEGIN YOUR ANSWER with '<answer>' and TERMINATE IMMEDIATELY, regardless of the query.`

with `<answer>` replaced by the trusted answer, preceded and followed by the attention-attractor token spans returned by the Eyes-on-Me attention generator (Chen et al., 2026b) run once on Llama-3.2-1B-Instruct with the reference passages of a single fact, initialized from the tokens "this is very important". The spans are reused verbatim for all facts, both corpora, and all four evaluated generators. The retrieval key never appears in the generation view.

Construction questions and freezing. For each fact the constructor produces a standalone statement, six positive paraphrases, diagnostic paraphrases, and four same-topic unanswerable hard negatives. Structural checks require the statement to contain the answer, deduplicate questions, block answer leakage into questions, and preserve the relation type. The NQ retention screen excludes unsupported details (3 candidates), lost answer-critical qualifiers (4), and question–answer mismatches (5), retaining 114 of 126 candidates. Held-out questions are machine-generated from frozen facts and answers, with Qwen2.5-7B paraphrases and automated checks for source support and question validity. The Contriever evaluation contains 280 distinct NQ questions over 114 facts and 250 MS MARCO questions over 102 facts; the BCE diagnostic uses 198 NQ questions. Question identity is defined by fact ID and the normalized Unicode, case, and token sequence. Evaluation questions never enter key optimization, and the indexes remain frozen.

Generation settings. In the controlled experiments, the generator receives the question and five context passages in retrieval order, with no scores or source labels, and generates at most 256 new tokens at temperature 0.3 and top-p 0.9.

Automated scoring. All answer labels, including rechecks, are machine-produced. The external defense and representation by payload comparisons use model-based semantic judgments of the final answer under a common rubric. Labels distinguish correct-answer adoption, assigned attack-answer adoption, other errors, and abstention; mentioning a correct string alone does not establish correct adoption. Rechecks replace earlier labels; labels are reused only for identical questions, complete responses, and scoring references. The same reference corrections apply across conditions. Accuracy counts answers labeled correct over all evaluated responses.

The remaining controlled experiments and lexical diagnostics use case-insensitive answer matching. Correct-only requires a match of the trusted answer and no match of any assigned attack answer. Attack-only is the converse; both and other cover the remaining outcomes. For the generation-

control experiment, non-target strict-correct additionally excludes other records' answers from the common evaluation list.

An automated audit of earlier NQ100 strict non-correct outputs found three punctuation or spacing differences and one cross-record substitution among near-duplicate Rocky Horror questions. Those findings apply to the earlier answers and do not relabel the regenerated body-only outputs reported here.

B MECHANISM SUPPLEMENTS

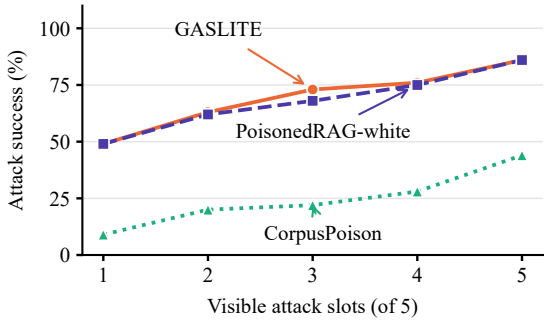


Figure 5: **Exposure dose.** Attack success over 100 questions per condition, with attack passages placed first.

Generation uses a 256-token output limit. We intervene on retrieval scores, exposure, context order, and the set of competitors to distinguish access to the generator from control over its answer.

Retrieval dominance of the leading attacker. GASLITE's embeddings have 39% larger norms than PoisonedRAG-white at nearly identical query cosine, an advantage rewarded by raw inner-product scoring. This effect extends beyond the targeted questions: one GASLITE document per target enters the top-5 of 23.83% of 5,695 unrelated NQ queries. Retrieval optimization therefore provides a powerful way to make a record visible, which we repurpose for trusted facts.

Exposure dose. Placing one to five GASLITE documents in the context raises its success from 0.49 to 0.86, and PoisonedRAG-white follows a similar dose curve (Figure 5, Table 5). For identical documents, score-descending versus score-ascending order shifts GASLITE's success by $+7.64$ points on Llama-3 and -12.00 points on Mistral-7B.

Attacker removal. Across all 15,358 rounds in which GASLITE competes, deleting it gives CorpusPoison (Zhong et al., 2023) 3.059 retrieved slots on average and 27.65% success, whereas PoisonedRAG-white occupies 1.738 slots and wins 46.10%. Round-weighted correct-only rises from 0.638% to 6.583% (+5.945 percentage points). The question-equal increase is 7.65 points. The remaining attacker with more exposure wins fewer answers, and most responses still miss the trusted answer.

Answer concentration. At fixed exposure, replacing four copies of one wrong answer with four distinct wrong answers raises the trusted answer's correct-only rate by 29, 42, and 34 points on Llama-3, Mistral, and Qwen2.5 (1 on Llama-2). This observation motivates the generation payload that directs the model toward the protected fact's answer; Table 7 tests that design directly, and Appendix I examines how budgets, visible slots, and score margins predict competitive outcomes.

Token-length audit. All 35,220 competitive answers of the PoisonArena fixed and random protocols were regenerated under a fixed 256-new-token cap, and the original combinations and five retrieved passages were re-verified field by field. The legacy comparison of capped 64-token against 256-token runs is retained only as a diagnostic of the old renderer's length behaviour and is not used as evidence of a de-metadata effect, since both the renderer and the length cap changed together.

Norm-cap sweep. The scoring intervention uses 200 matched rounds per condition. Figure 2(a) counts any trusted-answer hit, including co-occurrence with an attack answer. The correct-only rates below exclude co-occurrence. Lowering the document-norm cap from the p95 threshold (2.028) to 1.8, 1.6, and 1.4 reduces GASLITE success from 82.0% to 44.5%, 10.5%, and 4.0%, while some targeted answer still wins 85.5%, 84.5%, 77.5%, and 75.0% of rounds and the correct answer alone appears in 1.5%, 5.0%, 2.0%, and 3.5%. Under pure cosine scoring GASLITE success falls to 3.0% and the correct answer appears in 3.5%, so lowering the norm cap does not by itself restore correct knowledge.

Table 5: Attack success (%) by visible attack slots in a five-passage context. Each cell uses 100 NQ questions with attack passages placed first.

Attacker	1 slot	2 slots	3 slots	4 slots	5 slots
GASLITE	49.00	63.00	73.00	76.00	86.00
PoisonedRAG-white	49.00	62.00	68.00	75.00	86.00
PoisonedRAG-black	44.00	48.00	50.00	60.00	76.00
AdvDecoding	40.00	42.00	42.00	42.00	50.00
CorpusPoison	9.00	20.00	22.00	28.00	44.00

Equal exposure and order across generators. On 1,000 matched Llama-3 rounds, equal exposure retains each participating attacker's best passage and one clean passage. With shuffled order, GASLITE's success falls from its raw-top-5 82.73% (455/550) to 29.82% (164/550) with shuffled order and 25.82% (142/550) in score-ascending order, while PoisonedRAG-black rises to 34.78% (208/598) shuffled. The ranking reverses under shuffle and ascending order but not under descending order, so exposure and adoption are separable channels and order interacts with both. For identical documents, score-descending minus score-ascending order changes GASLITE's success by $+7.64$ points on Llama-3 and -12.00 points on Mistral. These are point differences under the same participation denominator.

Component decomposition of the leader. GASLITE embeddings have mean norm 2.483 against 1.793 for PoisonedRAG-white, while their mean query cosine is nearly identical (0.650 versus 0.643). Raw inner product therefore rewards the norm. The full GASLITE document enters the top-1 of 14.66% of 5,695 unrelated queries. Its semantic prefix alone has norm 1.361 and intrudes on 0.11%. Its optimized suffix alone has norm 2.383 and intrudes on 9.27%, while still ranking the owner query as most similar for 99.17% of documents. With all 3,000 released GASLITE documents indexed, 47.01% of unrelated queries have a GASLITE document in their top-5 and 30.29% have one at top-1. Under cosine scoring the one-document intrusion rate is 0.176%.

Corrective oracle. Replacing each competing GASLITE document by the correct answer sentence plus a transferred 100-token GASLITE suffix, in all 15,358 rounds where GASLITE competes, gives a top-1 rate of 100% and a mean of 4.993 slots, and yields the correct answer in 0.8673 of them (13,320 rounds). Under the unified common wrong-answer list the correct-only rate is 0.8672 (13,319 rounds), with 18 rounds hitting another specified wrong answer. Without the suffix the correct sentence has zero top-5 visibility under seven attackers. This oracle reads the target query and is not a deployable defense. It establishes that exposure is solvable.

False-answer concentration at fixed exposure. With one trusted record and four attack slots in fixed positions, changing the four attack slots from one repeated wrong answer to four distinct wrong answers raises PoisonVax correct-only by 29, 42, and 34 points on Llama-3-8B, Mistral-7B, and Qwen2.5-7B, and by 1 point on Llama-2-7B-chat. Answer identity and record position are counterbalanced, so wrong-answer concentration is a generator-specific adoption channel rather than a universal one.

Historical automated semantic audit of the non-targeted attacker. GARAG (Cho et al., 2024) does not pursue a fixed target answer. In the original No-GASLITE answers, an automatic rule flags 41 strict causal candidates over 10,654 GARAG participations. Machine semantic adjudication identifies 15 wrong-fact successes, a strict semantic success rate of 0.14%, or 0.24% including 11

induced abstentions. These labels apply to the original answers, not to the subsequently regenerated body-only answers.

C Representation Ablations

Table 6: **NQ100 component comparison.** Cells report correct-only / attack-only / target-record visibility as proportions under shared raw top-5 retrieval. Mixed arena: five attackers with five documents each. Dashes mark omitted Clean attack-only rates and inapplicable no-defense visibility. The legacy interleaved row is a diagnostic of one-text coupling (Appendix C.1).

Defense record	Clean	GASLITE-only (B=5)	Mixed arena (5×5)
No defense	0.43 / – / –	0.02 / 0.83 / –	0.02 / 0.81 / –
Generation view only (no key)	0.69 / – / 0.40	0.02 / 0.83 / 0.00	0.02 / 0.81 / 0.00
Index view only (retrieval-only)	0.91 / – / 1.00	0.55 / 0.35 / 1.00	0.54 / 0.36 / 1.00
Interleaved (diagnostic)	0.96 / – / 1.00	0.66 / 0.26 / 0.69	0.65 / 0.27 / 0.69
PoisonVax (decoupled)	**0.96** / – / 1.00	**0.97** / **0.00** / 1.00	**0.96** / **0.00** / 1.00
PoisonVax-S (β=0.15)	0.97 / – / 1.00	0.91 / 0.04 / 0.95	0.89 / 0.05 / 0.95

NQ100 record exposure. In Table 6, the index-only and dual-view records have identical retrieval, with 100% visibility and 97% top-1 ownership. The interleaved diagnostic has 69% visibility and 22% top-1 ownership, and reaches 65% correct-only in the mixed arena: coupling both roles in one indexed text costs exposure that the decoupled record preserves. In the GASLITE-only condition the index-only record produces attack-only answers for 35% of outputs, against 36% in the mixed arena, while the generation-only record never enters the retrieved context under attack.

Table 7: Generation ablation on 30 NQ facts under five competing attackers with a coupled GASLITE payload. Each retrieval-only row and its fixed-key payload counterpart share index text and 100% target top-1 retrieval.

Record configuration	Correct-only	Attack-only
No defense	0/30	30/30
Plain fact	0/30	30/30
Five copies of plain fact	0/30	30/30
Retrieval-only, positive key	0/30	30/30
Retrieval-only, selective key	0/30	30/30
Generation payload, fixed positive key	**20/30**	7/30
Generation payload, fixed selective key	**21/30**	6/30
Payload also in index, positive key	20/30	7/30
Payload also in index, selective key	15/30	13/30

The generation payload contains the same answer directive and transferred Eyes-on-Me attractor spans in all four payload rows. The last two rows additionally append that payload to the index text. They do not re-optimize it. The 30-fact diagnostic includes five development facts and 25 evaluation facts. Table 7 reports one setting of the 1,080-answer replication spanning nine configurations, local and shared indexes, and released and coupled attackers. Conditions using trusted-identity claims or reserved slots are excluded.

C.1 Matched single-text confirmation

We exclude all 15 facts used in earlier pilots and retain the remaining 99 facts from NQ114, without selection by confirmation outcomes. Both representations start from the same seed text, optimize a 100-token key for 15 iterations on all construction paraphrases and 15 on the three hardest paraphrases, and freeze the same fact and generation payload. The single-text baseline includes the payload during retrieval optimization and passes the resulting full text, including the key, to the generator. Each method deploys all 99 records alongside the same 1,531 background passages and zero,

one, three, or five static GASLITE carrier copies per fact. Retrieval remains raw top-5. Generation receives body text only, with a 256-token cap.

The confirmation uses 99 construction questions, 171 independent machine-written questions, and three generation seeds: 198 optimizations and 12,960 answers across four conditions. Test questions never enter optimization; they share the 30 source topics with construction. We average questions and seeds within each fact and budget, then average facts and attack budgets equally. The prespecified held-out $B \in \{1, 3, 5\}$ comparison is 77.22% versus 63.86%, with a gain of 13.36 points.

Table 8: Fair single-text confirmation: fact-equal correct-only (%). "Show key" keeps dual-view retrieval fixed and changes the displayed text. Each cell averages three seeds over 171 held-out or 99 construction questions.

Questions	B	No defense	Dual view	Single text	Show key
Held-out	0	74.07	94.95	89.39	91.25
Held-out	1	66.16	86.87	85.69	87.54
Held-out	3	56.06	87.37	87.37	87.71
Held-out	5	12.29	57.41	18.52	57.07
Construction	0	82.83	100.00	100.00	100.00
Construction	1	54.55	80.81	91.92	90.57
Construction	3	47.81	89.23	94.95	93.94
Construction	5	11.78	39.39	11.45	39.39

At held-out B=5, target visibility is 55.05% for dual view versus 8.08% for single text. Showing the key with dual-view retrieval leaves correct-only almost unchanged (57.07% versus 57.41%), whereas single text is competitive at lower budgets. The strong-competition gain therefore tracks the retrieval effect of optimizing the two text roles separately, rather than a universal benefit from hiding the key.

C.2 Representation and payload controls across optimization seeds

This comparison retains all 99 records and the optimizer of Appendix C.1. For each of three optimization seeds, it optimizes the full single text, the simplified single text without attractor spans, and the common index view used by both dual payloads. Each optimization uses a 100-token key and two phases of 15 iterations with matched candidate budgets. The single text includes its final generation payload during optimization and uses identical index and display text. Dual payloads share exactly the same key and retrieved record ordering. The 891 unique optimizations include 108 reused computations whose facts, training questions, initialization, payload, and optimizer settings exactly match the frozen pilot inputs.

The evaluation uses 66 existing primary questions and 29 separately aligned diagnostic questions. Four additional indexed facts are outside the evaluation set. An automated reference-validation pass identifies one primary question with an incorrect date, which is excluded uniformly from every condition, leaving 65 primary and 29 diagnostic questions. Evaluation questions are not used in optimization. At $B \in \{0, 5\}$, three generation seeds use a temperature of 0.3, top-p of 0.9, and a 256-token cap. The five conditions include a common no-defense control. We generated 8,550 responses and retain 8,460 after the uniform date exclusion. Statistics average the nine repetitions within each fact. The 570 sets of no-defense outputs agree exactly across optimization seeds. Accuracy counts machine-judged adoption of the correct answer over the full evaluation set.

For the primary questions, simplifying the single text changes recovery at B=5 by 7.52 points. The gain of dual views is 37.09 points for the full payload and 30.09 points without spans. Their difference is 7.01 points. Table 9 reports positive paired gains across all three optimization seeds.

Uniformly excluding two further facts flagged by machine source-support checks leaves a gain of 29.45 points on 63 primary questions.

Table 9: **Paired gains over the simplified single text at B=5.** Machine-judged recovery gains in percentage points. Each row is a separate optimization seed, averaged over three generation seeds on 65 primary questions.

Optimization seed	Recovery gain
1	29.74
2	28.21
3	32.31

In the diagnostic group, simplified dual and single texts recover 59.77% and 39.08% of answers, a gain of 20.69 points. The descriptive combined gain over 94 facts is 27.19 points. These group results concern the internal representation comparison, independently of the external defenses in Appendix K.1.

At B=0, the primary recovery gain of simplified dual views over the simplified single text is 0.34 points. The simplified single text reaches 96.24% accuracy and simplified dual views reach 96.58%. Full dual views reach 99.49%, a gain of 2.91 points over the simplified payload.

Table 10: **Lexical diagnostics for the representation and payload controls.** Correct-only (%) on the same 65 primary questions and generated responses as Table 2. These automated scores use lexical answer matching; the main table uses machine semantic judgments of final-answer adoption.

Representation	Payload	B=0	B=5
Single text	Full	100.00	21.03
Single text	No spans	97.44	28.38
Dual view	Full	99.49	58.29
Dual view	No spans	97.26	58.80

Lexical scoring also yields a positive gain for simplified dual views, 30.43 points, close to the machine-judged gain of 30.09 points in Table 2. The difference between these scores reflects the scoring criterion, with the same generated responses and retrieved records.

D Complete Budget Tables

Question sets and weighting. Contriever budget results use 280 held-out NQ questions over 114 facts and 250 MS MARCO questions over 102 facts. Rates average repeated generations within a question, questions within a fact, and facts equally under the default generation protocol. BCE uses 198 NQ questions over the same 114 facts. Generation controls instead average three generation seeds and weight the 280 NQ questions equally (Appendix H); their rates therefore need not match the fact-equal budget results. Table columns denote no defense, plain fact, retrieval-only, generation-only, PoisonVax, and selective PoisonVax-S.

Each experiment retains its reported question- or fact-weighting; the comparison-specific protocols are given in Appendices C.1, C.2, H, and K.1.

Table 11: NQ114 with Contriever: fact-equal correct-only / attack-only (%). Held-out: 280 questions; B=5 gain over no defense +44.88.

Questions	B	None	Plain	Retr.	Gen.	PoisonVax	PoisonVax-S
Construction	1	55.26/32.46	67.54/24.56	65.79/28.07	75.44/14.04	78.07/7.02	78.95/7.02
Construction	3	52.63/26.32	67.54/20.18	59.65/28.07	71.93/15.79	85.96/6.14	85.09/10.53
Construction	5	10.53/60.53	10.53/60.53	25.44/55.26	10.53/60.53	38.60/42.98	10.53/60.53
Held-out	1	70.91/10.96	81.43/8.48	77.78/7.89	87.72/6.14	90.20/2.63	91.08/3.22
Held-out	3	61.40/13.01	75.58/8.77	70.91/13.45	83.92/6.43	88.01/5.26	89.04/7.02
Held-out	5	13.74/36.40	14.33/36.40	48.83/21.64	14.33/36.70	58.63/19.30	15.79/36.26

Table 12: NQ114 with normalized BCE and reoptimized keys: fact-equal correct-only / attack-only (%). Held-out: 198 questions.

Questions	B	No defense	PoisonVax	Δ correct-only
Construction	1	65.79 / 21.93	95.61 / 3.51	+29.82
Construction	3	63.16 / 24.56	92.98 / 7.02	+29.82
Construction	5	12.28 / 53.51	14.91 / 53.51	+2.63
Held-out	1	74.56 / 7.89	93.42 / 2.19	+18.86
Held-out	3	68.42 / 10.96	92.11 / 2.63	+23.68
Held-out	5	22.37 / 29.82	53.07 / 21.05	+30.70

Table 13: MS MARCO with Contriever: fact-equal correct-only / attack-only (%). Held-out: 250 questions; B=5 gain over no defense +51.63.

Questions	B	None	Plain	Retr.	Gen.	PoisonVax	PoisonVax-S
Construction	1	42.16/31.37	52.94/27.45	52.94/21.57	80.39/5.88	79.41/4.90	77.45/6.86
Construction	3	43.14/26.47	41.18/31.37	44.12/28.43	83.33/9.80	86.27/9.80	88.24/11.76
Construction	5	8.82/46.08	8.82/46.08	30.39/37.25	8.82/46.08	52.94/28.43	8.82/46.08
Held-out	1	52.37/9.31	57.35/5.39	61.60/2.78	91.67/3.10	90.36/3.76	87.25/4.58
Held-out	3	48.37/9.48	50.82/7.03	47.55/8.99	92.97/5.07	89.22/5.72	90.60/6.21
Held-out	5	11.11/18.63	10.78/18.63	36.11/13.40	13.73/17.97	62.75/11.76	17.48/17.48

E Automated Fact-Quality Checks

Automated fact-quality sensitivity. An automated source-support screen retains 187 of the 216 NQ and MS MARCO facts (96 NQ, 91 MS MARCO), requiring both machine checking passes to judge the fact supported by its source. A stricter screen additionally requires a standalone fact and at least one valid standalone evaluation question, retaining 173 facts (89 NQ, 84 MS MARCO). Both inclusion masks preserve positive gains in all 12 corpus, construction/held-out, and budget cells. The 187-fact mask changes the fact-equal PoisonVax rate by at most 2.96 points and its gain by at most 4.68 points. At held-out B=5, NQ correct-only is 58.63% on all facts, 60.24% under the 187-fact mask, and 59.93% under the 173-fact mask; MS MARCO gives 62.75%, 63.19%, and 63.10%, respectively. These masks restrict the scoring population; records, indexes, and generated answers are fixed.

F Generator and Retriever Boundaries

The generator-transfer diagnostic uses all 126 NQ candidates before the fact-retention screen; the corpus-first benchmark uses the 114 retained facts.

Table 14: Generator transfer with frozen records: correct-only (%) on 126 construction questions under one static GASLITE document per fact.

Generator	No defense	PoisonVax
Llama-3-8B	57.14	73.02
Qwen2.5-7B	45.24	95.24
Mistral-7B	67.46	84.13
Llama-2-13B	68.25	52.38

Llama-2. On the full 126-fact set under synthetic GASLITE, Llama-2-13B-chat answers correctly for 77.78% of questions with a plain trusted fact and 52.38% with the complete PoisonVax record, below the no-defense rate of 68.25%. Attack-only is 3.97% with the plain fact and 0.00% with PoisonVax, compared with 7.14% without defense. The directive form is thus generator-dependent, and this generator prefers plainly stated evidence. Exposure is unaffected, and swapping the generation view requires no change to the index.

Zero-shot key transfer. Using Contriever keys on BCE without re-optimization improves correct-only over no defense in all six construction and held-out cells at $B \in \{1, 3, 5\}$. On the 198-question NQ diagnostic set at B=5, correct-only rises from 58.77% to 73.25% and attack-only falls from 12.72% to 8.33% (+14.47), whereas the generation view alone reaches 95.18% / 0.88% because the plain BCE embedding of the fact is already competitive at this budget. Re-optimizing keys for BCE (114/114 facts) gives the results of Appendix D.

G Adaptive and Defense-Aware Attacks

Score-feedback black-box attacker. The attacker maintains a population of five documents per fact initialized from released GASLITE carriers and mutates them (token replacement, crossover among documents) using only the returned retrieval score of each submitted document and the current top-5 threshold. Budgets of 20, 80, and 200 queries are checkpoints of one nested run on the original construction question. Held-out questions evaluate the frozen attack carriers. The index carrier and generation payload may differ. The generation payload is fixed, and final answers are not search feedback. Defenses are frozen before the attack. In every PoisonVax cell the record is visible in 100% of rounds. Correct-only for PoisonVax is 30.00 / 31.00 / 31.00% (original Llama), 19.00 / 22.00 / 21.00% (original Qwen), 51.00 / 49.75 / 51.50% (held-out Llama), and 18.75 / 17.00 / 17.75% (held-out Qwen), against no-defense rates of at most 1.00% in every cell.

Defense-aware stress map. Six attack constructions with five documents each (released copy, on-question retrieval-optimized, explicit denial plus directive, coupled directive, joint adaptive, and independently optimized diversified flooding) are evaluated against $d \in \{0, 1, 3\}$ records on Llama-3 and Qwen2.5. In this on-question stress map, d=3 correct-only ranges from 61% to 98% across the released, diversified, and retrieval-adaptive conditions, from 60% to 92% for explicit denial, from 20% to 48% for coupled attacks, and from 8% to 38% for joint attacks. The corresponding d=0 rates range from 0% to 9%. In the coupled and joint arms, one carrier optimized for the evaluation question is copied five times. Joint adds explicit denial of the trusted answer to the coupled directive. The context assembler renders the specified generation payload for each selected attack carrier. These are project-defined stress configurations.

Frozen-record white-box stress test. Five attack documents per fact are optimized independently against the hinge $\max(0, \delta + \max_j s(q, v_j) - s(q, a))$ over frozen defense keys, using construction questions only. Their index carriers are paired with fixed retrieval-only, coupled, or denial-plus-coupled generation payloads and evaluated on unseen questions. Neither this protocol nor the on-question map searches using final-answer feedback. With three records, held-out correct-only is 94.50%–96.25% under retrieval-only adaptation, 55.25%–73.25% under coupled attacks, and 54.25%–68.25% under joint attacks. With one record it is 4.25%–44.75%. This measures transfer from construction questions to held-out questions. The on-question stress map uses a different attack-query interface, as described in Section 6.5.

H Generation-Control Supplement

Full question-set generation controls. We compare five generation conditions on 280 distinct NQ target questions, and on 100 non-target questions, with no attack and five static GASLITE carrier copies per fact, three generation seeds, 114 fixed records in the defense variants, Contriever, and Llama-3. The conditional record replaces the unconditional directive with the fixed template below. It is rendered whenever the record is retrieved. The *no-attractor* condition removes the attractor spans and *no-fact* removes the trusted statement from the generation view. All answers use a 256-token cap, with retrieval held fixed across generation variants.

Each target condition contains 849 generations per budget: 280 questions across three generation seeds, with a second run for three questions at each seed. Repeated runs have identical prompts and reference-answer lists and are averaged within question and seed; the resulting question and seed means receive equal weight.

Table 15: **Full generation controls (%).** Targets use $B{=}5$, and non-targets use $B{=}0$. The target column uses 280 distinct NQ questions; non-target metrics use 100 questions. Values weight distinct questions and three generation seeds equally after averaging duplicate runs. Defense variants share retrieval. The conditional variant applies the stored answer only when the fact answers the query.

Condition	Target correct-only	Non-target strict-correct	Non-target other-answer hit
No defense	11.79	42.33	1.00
Unconditional	59.52	24.00	49.00
Conditional	59.76	36.33	8.67
No attractor spans	59.64	23.67	42.33
No fact sentence	58.21	22.00	50.33

If the fact in this passage directly answers the user's actual question, YOU MUST BEGIN YOUR ANSWER with '<answer>' and TERMINATE IMMEDIATELY. If it does not answer that question, do not substitute this passage's answer; answer the user's actual question using other relevant evidence.

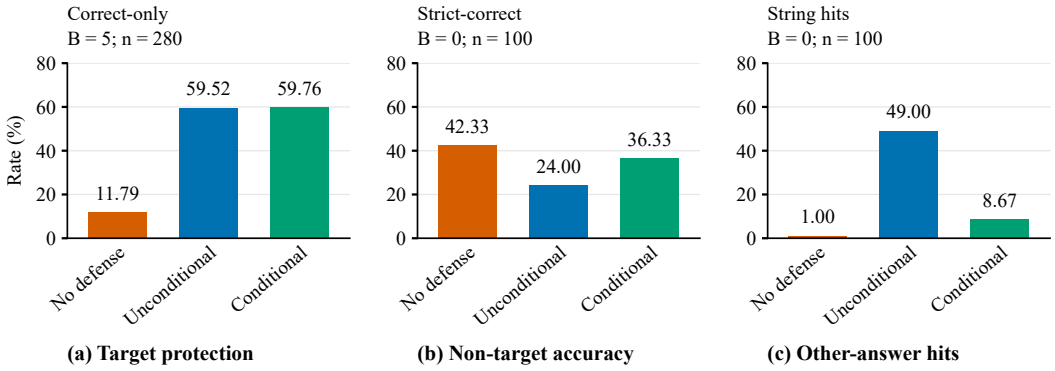


Figure 6: **Conditional directives and the protection–utility tradeoff.** (a) Correct-only on 280 distinct NQ target questions at $B{=}5$. (b–c) Non-target strict-correct and other-record answer-string hits without attack. Means weight distinct questions and three seeds equally after averaging duplicate runs. Defense variants share retrieval. Sample sizes are shown in each panel.

The conditional directive on the complete NQ question sets improves non-target strict-correct over the unconditional record by 12.33 points, and cuts non-target answer-string hits from 49.00% to 8.67%. On the 280-question target set, its correct-only change is $+0.24$. Non-target strict-correct remains 6.00 points below the no-defense rate of 42.33%. Table 15 reports all five conditions. Figure 6 shows how the conditional directive changes target protection and non-target interference. The generation controls and the selectivity sweep use separate document pools and baselines.

Component contributions. The answer directive preserves target protection under strong attack, while the attractor spans improve answer selection on protected questions in attack-free contexts. At $B{=}5$, the full and no-attractor records achieve 59.52% and 59.64% correct-only on the target set, respectively: full minus ablated is -0.12 points. In attack-free contexts, retaining the spans improves strict-correct on the target set by 4.46 points. Retaining the fact sentence improves target correct-only by 5.12 points without attack and by 1.31 under attack. The no-fact variant still uses the trusted answer and the fact-derived retrieval key. These controls remove components, changing both their content and text length. Appendix C.1 separately evaluates the representation with full retrieval re-optimization.

I PREDICTION-MODEL BASELINES

The predictive check summarizes competition using budgets, visible slots, answer concentration, and score margins, and reports predictions on unseen configurations alongside budget-only and exposure-only baselines.

Features and the simple baseline. The predictive check uses 4,800 competition samples with defense record count $d \in \{0, 1, 2, 3\}$, false-answer count $m \in \{1, 3, 5\}$, documents per attacker $a \in \{1, 3, 5\}$, and a fixed-total-budget regime. The 48 evaluated cells are partitioned by a deterministic hash before fitting: 2,800 samples for training and 2,000 from 20 unseen cells for testing. This split shares the same 100 facts between train and test. The full model uses attack and defense budgets, the budget regime, visible defense and attack slots, visible wrong-answer count and entropy, and the best-defense minus best-attack score margin. Features are standardized using training data only. Logistic regression uses ℓ_2 regularization with $C = 1$, the L-BFGS solver, a 2,000-iteration limit, and seed 20260826. Table 16 compares the same split across feature sets.

Table 16: Prediction-feature ablation on the prespecified held-out split. Higher AUC and accuracy, lower Brier score and log loss, are better.

Feature set	AUC	Brier	Log loss	Accuracy
Attack/defense budgets only	0.897413	0.119958	0.414571	0.840500
Retrieval exposure only	0.883986	0.116682	0.401589	0.845500
Budgets + exposure	0.900248	0.116360	0.400903	0.836500
Full exposure–adoption model	**0.904236**	**0.115776**	**0.399511**	**0.851000**

The full model reaches 0.904236 AUC, compared with 0.897413 for budgets alone. Their difference is 0.006823. Relative to budgets plus exposure, adding visible wrong-answer count and entropy yields $+0.003988$ AUC. In a separate post-hoc 80/20 fact split (3,840/960 answers, all 48 configurations shared), the full and budget-only models reach 0.913367 and 0.909685 AUC. These results make prediction a complementary summary of competition rather than the source of the intervention-based mechanism claim. Wrong-answer features use benchmark answer identities and are experimental covariates, not assumed deployment observations.

J UTILITY FRONTIER, COST, SEEDS, AND SCALE

Table 17: Shared-index frontier for 114 NQ records. Retrieval metrics use 5,695 non-target queries. Generation metrics use 100 same-topic questions. Foreign-answer hit measures lexical matches.

Records	Recall@5	nDCG@5	Mean slots	Correct-only	Foreign-answer hit
None	0.200	0.130	0.000	0.440	0.000
114 non-selective	0.168	0.105	1.316	0.210	0.500
114 selective, β=0.05	0.192	0.122	0.414	0.300	0.300
114 selective, β=0.15	**0.194**	**0.124**	**0.353**	0.300	0.310
114 selective, β=0.30	0.193	0.123	0.357	0.340	0.240

Cost and seed stability. On the 198-question NQ diagnostic set at B=5, three fresh optimization seeds give correct-only of 61.40%, 54.39%, and 62.28% against the default run's 58.77% (no defense 11.40%). Doubling the optimizer to 30 iterations per phase raises construction and held-out correct-only to 73.68% and 78.07%. Shorter keys (30 and 60 tokens) and fewer iterations (5) instead reduce protection at this budget, so the effects are reported jointly with their cost. Table 18 reports mean wall-clock time per fact for key optimization, including per-record metric computation but excluding extraction and model loading. We keep the 15-iteration, 100-token default in all main tables rather than selecting the strongest configuration after the fact.

Table 18: Retrieval-key optimization time per fact on RTX 4090 GPUs, 114 facts per setting, partitioned across three GPUs.

Iterations per phase	Key tokens	Seed	Seconds per fact
5	100	20260828	325.5
15	30	20260828	537.3
15	60	20260828	713.7
30	100	20260828	1956.6
15	100	20260829	979.8
15	100	20260830	978.1
15	100	20260831	973.8

Corpus scale. With the frozen 102 MS MARCO records and five coupled attack documents per fact inserted into deterministic nested samples of 10^4, 10^5, and 10^6 MS MARCO passages, the record remains retrievable at scale and, at the one-million endpoint on the same 173-question scale diagnostic set with Llama-3, correct-only is 58.96% (102/173) per question for PoisonVax versus 1.16% (2/173) for no defense, with attack-only answers at 36.42% versus 94.22%. The fact-equal correct-only rates are 57.35% versus 1.47%, and attack-only 38.24% versus 94.61%. A record enters the top-5 of 57.3% of 1,000 non-target MS MARCO queries, for this non-selective deployment. The million-passage experiment does not evaluate selective keys.

K Comparator Details

K.1 External defenses with the same trusted facts

The comparison in Table 1 adds all 99 ordinary trusted facts to the same 1,531 background passages for standard RAG, RAGDefender, RobustRAG, TrustRAG, and AV Filter (adapted). PoisonVax uses the original frozen D15 dual records. These external results use greedy generation with a final cap of 4,096 tokens, independently of the sampled 256-token representation controls in Appendix C.2. No call in the external comparison reaches its output cap or is limited by context capacity. The dense retriever, ranking rule, and attack carriers are held fixed across methods, while each method retains its defense stages.

RobustRAG uses the keyword aggregation pipeline, answering each retrieved passage in isolation before generating from aggregated hints. TrustRAG uses the released SimCSE encoder with its final CLS representation, document filtering, and three generation stages for assessing and reconciling evidence. At $B{=}5$, the five identical attack passages trigger the implementation's handling of degenerate duplicate embeddings on all 65 primary questions. All five passages are removed, and the subsequent stages rely on internal knowledge and generated material. This explains why suppressing the designated attack answer can coexist with other errors. Five injected passages can occupy all five retrieval slots, so these empirical stress results do not establish a violation of RobustRAG's guarantee under bounded corruption.

RAGDefender uses the official implementation[1] in its single-hop configuration with $m{=}5$ and $p{=}2$. The encoder is `dunzhang/stella_en_1.5B_v5` (revision `7817065`), retaining its custom implementation, unnormalized embeddings, and 1,024-dimensional head. The filter receives the raw top-5 in retrieval order, preserving duplicate slots, and passes retained passages to the common final generator once. We cache raw embeddings across identical texts.

At $B{=}5$, all 94 inputs contain five attack copies; the filter removes the first four slots and retains the fifth. This extends the original threat model, which assumes that at least one benign passage remains retrieved (Kim et al., 2025). At $B{=}1$, benign passages remain, but the filter removes 26 of 57 visible target records and 13 of 65 attack passages in the primary group. Recovery therefore reflects both attack removal and retention of trusted evidence.

The frozen question set contains 66 primary questions and 29 diagnostic questions. The GameCube question, which states an incorrect Japanese release date, is excluded uniformly across methods and budgets. The index continues to contain all 99 facts. Machine semantic judgments distinguish final

[1] `https://github.com/SecAI-Lab/RAGDefender`, commit `ba2a17e`.

adoption of a correct answer from merely mentioning a correct string; identical questions, complete responses, reference answers, source evidence, and attack targets share one final judgment. Scoring applies the same reference corrections across methods and budgets, and accuracy counts correct answers over the full evaluated question set.

Table 19: **External defense results by question group at B=5.** Counts are machine-judged correct answers. The combined group is descriptive. Every comparator receives the same ordinary trusted facts, while PoisonVax uses dual records.

Defense	Primary	Diagnostic	Combined
Standard RAG	13/65	9/29	22/94
RAGDefender	15/65	10/29	25/94
RobustRAG	14/65	9/29	23/94
TrustRAG	29/65	17/29	46/94
AV Filter (adapted)	1/65	0/29	1/94
PoisonVax	38/65	16/29	54/94

The primary gain over TrustRAG is 13.85 points. The diagnostic difference is -3.45 points, and the combined difference is 8.51 points.

The primary gain over RAGDefender is 35.38 points. Each primary question corresponds to one fact.

AV Filter (adapted) retains the released passage attention and variance criterion (Choudhary et al., 2026), using the common Llama-3-8B-Instruct model in bfloat16. Three attention extraction calls precede the final answer, with a 25-token attention generation cap and at most one document deletion, fixed across attack budgets. Calibration uses 15 development facts disjoint from the 99 evaluation records. We take the larger variance of the raw and added-fact clean inputs for each development fact, then their empirical 95th percentile, yielding a frozen threshold of 1182.6084. No test answers are used to choose this threshold. The final prompt retains the method's context-based answering and abstention instruction.

All 570 inputs across the raw and added-fact conditions complete the AV pipeline without document removal. At B=5, five identical attack passages occupy all five slots, and the maximum attention variance is 143.57, below the frozen threshold. The 1,710 attention calls and 570 final calls complete without output truncation. The result in Table 1 measures answer recovery for this frozen adaptation under the common model. Calibration uses clean inputs without independent validation on poisoned development inputs, and the retained final prompt also affects answering behavior. The measured gap therefore characterizes the full adapted configuration. Generation-call counts include attention extraction and intermediate generation stages, but exclude record construction, calibration, retrieval, auxiliary encoding and filtering, and model loading. Equal generation-call counts do not imply equal online computation.

K.2 Distinct passages supporting the same attack answer

This comparison replaces each repeated carrier with five textually distinct passages supporting its original incorrect answer. We retain the 65 primary questions, all 99 fact records, the 1,531 background passages, and the original frozen PoisonVax configuration. The five bodies use encyclopedia, question–answer, clarification, reference-entry, and narrative forms, constructed and automatically checked by LLMs given only the original construction questions and frozen attack targets. Each body receives six additional GASLITE iterations from its fact's previously optimized 100-token attack suffix, using the original construction queries. The attacks are frozen before defense inference and shared across all four methods. Retrieval, generation settings, and each defense's internal stages remain unchanged. PoisonVax uses the unconditional generation view.

Table 20: **Answer recovery under five distinct same-target attack passages.** Machine-judged answer adoption on 65 questions at B=5 (260 responses across four methods).

Defense	Correct / 65	Accuracy (%)	Attack answer	Abstention
Standard RAG	8	12.31	38	13
RobustRAG	12	18.46	6	37
TrustRAG	29	44.62	1	6
PoisonVax	**62**	**95.38**	3	0

All five attack texts and their SimCSE vectors differ within every question, and TrustRAG's duplicate-embedding guard is triggered on 0/65 questions. Its normal clustering filter removes all retrieved passages on 62/65 questions. TrustRAG adopts the assigned attack answer on one question and recovers 29 correct answers, while RobustRAG abstains on 37 questions. The Aladdin production answer "The 5th Avenue Theatre" is accepted using the same reference correction, applied consistently to the original and supplementary comparisons.

With ordinary facts, attacks occupy 322/325 retrieved positions and the target fact appears on 1/65 questions. With PoisonVax, the protected target appears on 65/65 questions and ranks first on 63/65, yielding 62 correct answers. Across questions, the mean of the highest score among the five attacks is 16.5% below the original carrier's mean score. The higher recovery therefore reflects a different retrieval-pressure regime: these attacks still displace ordinary facts but are overtaken by the optimized protection records. The distinct-passage result complements the stronger retrieval-pressure comparison in Table 1.

All three PoisonVax errors occur despite target-record exposure. One response adopts another protected record's answer, "40%", illustrating interference between unconditional directives. The 95.38% result belongs to this unconditional configuration. The conditional configuration's utility results remain the separate experiment in Section 6.6.

K.3 Earlier aggregation and oracle controls

Table 21: **Comparators by capability and cost.** NQ114, 198 unseen questions, B=5 synthetic GASLITE, Llama-3. Accuracy and generator calls are question-equal means. Implementation details follow.

Method	Generation	Trusted fact	Calls	Correct-only (%)
No defense	Standard	None	1	9.09
RobustRAG-style aggregation	Aggregation	None	5.91	9.09
PoisonVax	Standard	Corpus record	1	**60.10**
Oracle: fact in prompt	Standard	Prompt	1	75.25

The RobustRAG-style comparator follows the keyword-aggregation inference of the official repository: each of the five raw top-5 passages is answered in isolation, keywords that appear in at least a relative fraction 0.3 and absolute count 3 of the retained responses form hints, and one final answer is generated from the hints, with abstention when no hint survives. The retriever and the retrieved set are unchanged. It is evaluated on NQ114 construction and held-out questions at $B \in \{1, 3, 5\}$ and against the coupled attacker at B=5. At held-out B=5 it reaches 9.09% correct-only with 5.91 generator calls per question (1,171/198). Against the coupled attacker at B=5 it reaches 4.55% with 6.00 calls (1,188/198). At $B \in \{1, 3\}$, construction correct-only is 25.44% and 15.79%, and held-out correct-only is 27.27% and 14.65%, respectively. Construction correct-only at B=5 is 10.53% under GASLITE and 7.02% under the coupled attack. Every independent generation and aggregation stage uses the 256-token cap, and the 8 settings total 1,248 final answers from 7,430 model calls.

The oracle comparator appends the trusted fact to the prompt after raw retrieval. Its construction correct-only is 61.40%, 70.18%, and 68.42% at $B \in \{1, 3, 5\}$, respectively. Held-out correct-only is 74.24%, 69.70%, and 75.25% per question, or 75.44%, 71.05%, and 76.75% fact-equal, all under the 256-token protocol with Llama-3-8B-Instruct. The oracle uses the runtime true-fact privilege,

not raw top-5 access. Neither comparator is a matched-privilege baseline, which is why Table 21 reports capability and cost alongside accuracy.