IMPLICIT RANKING LEAKAGE IN BLACK-BOX RAG: RETRIEVAL STABILITY-GUIDED KNOWLEDGE CORRUPTION ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-Augmented Generation (RAG) is widely used in large language model systems. Recent studies show that RAG is vulnerable to knowledge corruption attacks (KCAs), where adversaries observe top-k retrieved documents to approximate the retriever and manipulate downstream generation. Existing attacks mainly use top-k membership as binary supervision and treat all retrieved documents as equally important. This ignores hidden ranking information and underestimates the risk exposed by retrieval behavior. In this paper, we propose Ranking Vulnerability-enhanced Knowledge Corruption Attack (RV-KCA), a black-box method that infers relative document retrieval stability from top-k membership and uses it as pairwise supervision to conduct stronger knowledge corruption attacks. First, RV-KCA applies semantics-preserving query perturbations and estimates retrieval stability from how often each document remains in the top-k set. Our theoretical analysis shows a monotonic relationship between retention probability and the original retrieval score under a shared local noise model. It also characterizes sampling uncertainty in the stability estimates, which serve as noisy evidence of relative document ordering. Second, RV-KCA selects document pairs with clear stability differences and uses them to train a margin-aware surrogate retriever. The surrogate model learns to assign higher scores to more stable documents while preserving membership supervision information derived from the observed top-k sets. Finally, RV-KCA optimizes query-side triggers on the surrogate to give the injected document a ranking advantage over estimated boundary competitors across query variations. Experiments show that RV-KCA achieves higher surrogate fidelity and overall attack success rates than existing black-box baselines. It also reaches a fixed surrogate fidelity target with fewer target-retriever calls. These results demonstrate that retrieval stability exposes useful ranking information for knowledge corruption attacks even when similarity scores and explicit ranks are hidden.

1 INTRODUCTION

Retrieval-Augmented Generation (RAG) has become a standard approach to extending large language models with external knowledge and is widely used in high-stakes applications. Grounding responses in retrieved documents can improve factual accuracy Oche et al. (2025); Wang et al. (2024). However, this reliance on external evidence also creates a security risk because corrupted documents can influence the generated responses Yu et al. (2024); Xue et al. (2024); Gulyamov et al. (2026). Recent studies show that adversaries can use black-box retrieval outputs to approximate the retriever and induce attacker-desired responses through knowledge corruption attacks (KCAs) Bagwe et al. (2025); Zou et al. (2025). These findings show that assessing RAG security requires understanding both what retrieval outputs reveal and how attackers can exploit that information.

Existing black-box KCAs promote malicious documents so that the generator uses them as evidence, as illustrated in Fig. 1. Zou et al. (2025) inject documents designed to match target queries and steer generation toward attacker-specified answers. Chen et al. (2025) train a surrogate retriever from observed top-k query-document interactions and use it to optimize adversarial triggers that promote target documents. Such retriever approximation mainly relies on whether a document appears in 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.

top-k set Xiang et al. (2024); Chen et al. (2025); Naseh et al. (2025). Existing research overlooks the ranking information revealed by changes in "Top-k membership," thereby underestimating the threat such information leakage poses to black-box RAG systems.

However, existing methods treat all retrieved documents as equally important and use them to supervise the surrogate retriever training Wang et al. (2025b); Li et al. (2025b); Cheng et al. (2024). This overlooks hidden ranking information that may still be inferred from retrieval behavior Pang et al. (2025); More et al. (2025). Recovering this information could help attackers identify which documents the target retriever ranks more highly and use their relative ordering to construct more informative supervision. Such supervision could help train a faithful surrogate with fewer black-box interactions and enable more efficient KCAs. Ignoring these ranking differences therefore underestimates the threat posed by retrieval behavior. Thus, our central research question is: under what conditions retrieval stability provides statistically reliable information about hidden document ordering from limited top-k observations and how this information can guide surrogate training for stronger black-box knowledge corruption attacks.

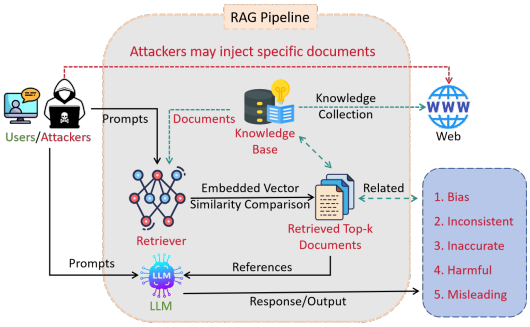


Figure 1: Security risks in black-box RAG systems: in a standard RAG pipeline, attackers can inject crafted documents into the external knowledge base. These documents can bias the LLM's generation, leading to inaccurate or harmful outputs.

To answer this question, we propose Ranking Vulnerability-enhanced Knowledge Corruption Attack (RV-KCA), which infers relative document ordering from retrieval stability and uses this ordering to train a surrogate retriever for stronger black-box knowledge corruption attacks. RV-KCA first applies controlled, semantics-preserving query perturbations and estimates retrieval stability from how often each document remains in the top-k set. The intuition is that documents farther above the retrieval cutoff can tolerate larger score changes before leaving the retrieved set Wang et al. (2025a); Gwak et al. (2022). Our theoretical analysis formalizes this connection by showing that retention probability is non-decreasing in the original retrieval margin under a shared local noise model. We also bound the error in estimating retention probability from a finite number of perturbations. This explains when stability can reveal hidden ordering and quantifies the uncertainty in its estimation. Then, RV-KCA selects document pairs with clear stability differences and converts them into pairwise supervision for a margin-aware surrogate retriever. The surrogate learns which documents the target retrieves and assigns higher scores to the more stable document in each selected pair. Finally, RV-KCA optimizes query-side triggers on the surrogate to promote injected documents across local query variations. By exploiting ranking information in repeated top-k responses, RV-KCA improves surrogate fidelity and enables stronger knowledge corruption attacks. It also reaches a fixed surrogate fidelity target with fewer target-retriever calls.

We evaluate RV-KCA on two RAG benchmarks across multiple retrievers and LLMs. RV-KCA achieves higher surrogate fidelity and overall attack success rates than existing black-box baselines under limited query budgets. It also reaches a fixed fidelity target with fewer target-retriever calls. These results show that retrieval stability exposes exploitable ranking information in top-k responses even when similarity scores and explicit ranks are hidden. In summary, our contributions are as follows:

- We propose Ranking Vulnerability-enhanced Knowledge Corruption Attack (RV-KCA), which estimates retrieval stability from unordered top-k responses under query perturbations. Our analysis establishes a monotonic relationship between retention probability and retrieval margin under a shared local noise model and also characterizes sampling uncertainty under independent sampling.
- We develop a surrogate training method that uses margin-based pairwise ranking to learn document preferences inferred from retrieval stability. The surrogate guides query-side trigger optimization against estimated boundary competitors across query variations.

- We evaluate RV-KCA on two RAG benchmarks across multiple retrievers and LLMs. RV-KCA achieves higher surrogate fidelity and overall attack success rates than existing black-box baselines under limited query budgets. It also reaches a fixed surrogate fidelity target with fewer target-retriever calls.

2 Related Work

Existing attacks show that malicious documents can influence the responses of RAG systems Bagwe et al. (2025); Xue et al. (2024). Zou et al. (2025) inject documents designed to match target queries and steer generation toward attacker-specified answers. Other studies explore how adversarial inputs affect retrieval and downstream generation Chaudhari et al. (2024); Cho et al. (2024); Li et al. (2025a). Closely related attacks approximate the target retriever from observed top-k outputs Chen et al. (2025); Naseh et al. (2025). Chen et al. (2025) use the resulting surrogate to optimize adversarial triggers that promote target documents. Such retriever approximation mainly relies on membership labels, which provide no explicit supervision for ordering retrieved documents Naseh et al. (2025). This overlooks hidden ranking information that could support more effective attacks. Our theoretical analysis explains when retrieval stability under query perturbations can reveal this ordering. RV-KCA converts retrieval stability into pairwise ranking supervision and achieves higher surrogate fidelity and attack success rates than existing black-box baselines under limited query budgets. Further discussion appears in Appendix A.

3 Proposed RV-KCA

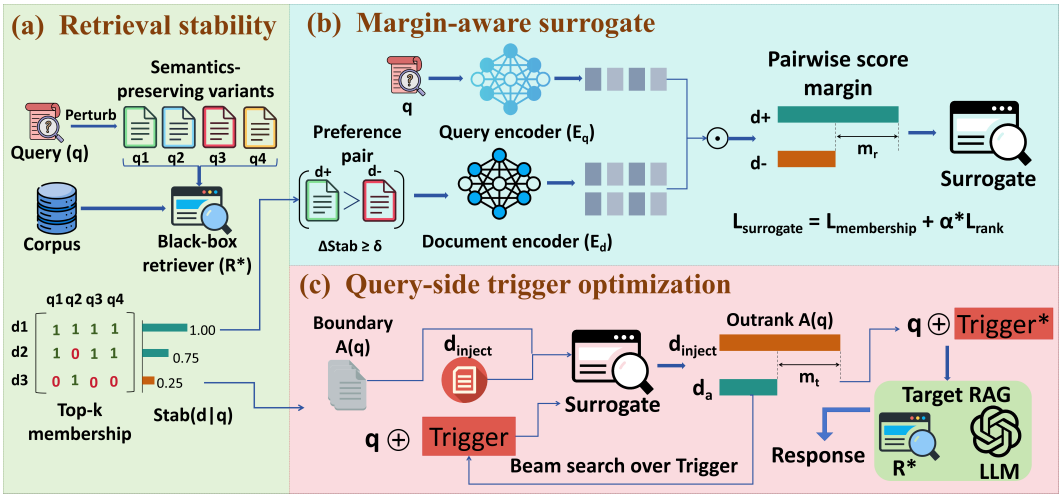


Figure 2: RV-KCA converts retrieval stability under semantics-preserving query perturbations into pairwise supervision for a margin-aware surrogate. The surrogate guides query-trigger optimization to promote injected documents over estimated boundary competitors.

3.1 Threat Model and Overview

As shown in Figure 1, we consider a RAG system with a target retriever, an external corpus $\mathcal{C}$, and an LLM. For a query q, the attacker observes only the unordered document set $D_k(q)$ containing the k highest-scoring documents. The target score $s_\star(q, d)$ and the ordering within this set are hidden. The attacker has no access to target parameters or gradients. The attacker can inject a limited number of controlled documents and submit queries augmented with a short query-side trigger.

RV-KCA aims to promote an injected document so that its content influences the generated response. It first estimates retrieval stability from repeated top-k membership observations under semantics-preserving query perturbations. It then converts stability differences into pairwise supervision for a margin-aware surrogate retriever. Finally, the surrogate guides query-side trigger optimization

against estimated boundary competitors. The analysis below states when stability differences reveal hidden document ordering and when their empirical estimates support reliable comparisons.

3.2 Retrieval Stability and Hidden Document Ordering

For each query q, let $\mathcal{P}_\lambda(q)$ be a fixed distribution of accepted query perturbations, where λ controls perturbation strength. We use lexical substitution and structural rephrasing, together with neutral contextual expansion and suppression of nonessential anchor terms. Candidates are filtered to preserve the original retrieval intent before target queries are issued. The corpus remains fixed throughout probing. For $d \in D_k(q)$, define its retention probability and empirical retrieval stability as

$$p_d(q) = \Pr_{q' \sim \mathcal{P}_\lambda(q)} \left(d \in D_k(q')\right), \quad \widehat{p}_d(q) = \frac{1}{m} \sum_{t=1}^{m} \mathbf{1}\left\{ d \in D_k(q^{(t)}) \right\}. \tag{1}$$

Here m is the number of accepted perturbations submitted to the target. Including the original query, probing costs $m + 1$ target-retriever calls. Stability is a query-conditioned retention frequency and does not estimate an absolute similarity score.

To connect stability with hidden ordering, let $\tau_k(q)$ be the k-th largest target score and define the retrieval margin $\gamma_d(q) = s_\star(q, d) - \tau_k(q)$. Documents with larger margins can tolerate larger adverse score changes before leaving the retrieved set. Appendix B.1.1 formalizes this sufficient retention condition. The following theorem gives a sufficient condition for inferring the original document ordering from retention probabilities.

Theorem 1 (Retention probability and retrieval margin). Fix q and suppose that no documents tie in score at the retrieval cutoff for q or for sampled perturbations, almost surely. Write the perturbed margin as

$$s_\star(q', d) - \tau_k(q') = \gamma_d(q) + Z_d(q'). \tag{2}$$

Under the shared local noise model, $Z_d(q')$ has the same marginal distribution across the compared documents. If Z follows this distribution, then

$$p_d(q) = \Pr(Z \geq -\gamma_d(q)). \tag{3}$$

Thus, retention probability is non-decreasing in the original retrieval margin. In particular, $p_{d_i}(q) > p_{d_j}(q)$ implies $s_\star(q, d_i) > s_\star(q, d_j)$.

The noise term includes changes in both the document score and the retrieval cutoff. Its shared distribution is an analysis assumption and does not follow from semantic preservation alone. The theorem allows dependence between documents within one retrieved set. It provides a partial ordering because different margins can have the same retention probability. Appendix B.1.2 gives the proof.

Theorem 2 (Finite-sample reliability). For m independent perturbations from the fixed distribution $\mathcal{P}_\lambda(q)$ and any failure probability $\eta \in (0, 1)$, with probability at least $1 - \eta$,

$$\max_{d \in D_k(q)} |\widehat{p}_d(q) - p_d(q)| \leq \varepsilon_m(\eta), \qquad \varepsilon_m(\eta) = \sqrt{\frac{\log(2k/\eta)}{2m}}. \tag{4}$$

On this event, every pair with $\widehat{p}_{d_i}(q) - \widehat{p}_{d_j}(q) > 2\varepsilon_m(\eta)$ has $p_{d_i}(q) > p_{d_j}(q)$. Under Theorem 1, each such pair also follows the original target-score ordering.

This result follows from Hoeffding's inequality and a union bound. It separates sampling uncertainty from the assumption connecting retention to ordering. The same bound holds under independent sampling with a fixed allocation across perturbation families, using the allocation-weighted retention probability. The proof and this extension appear in Appendices B.1.3 and B.1.4. The quality of the empirical ranking signal at the probing budgets used by RV-KCA is evaluated experimentally.

3.3 Margin-Aware Surrogate Training

RV-KCA retains a document preference when its empirical stability gap is at least a fixed empirical resolution threshold $\delta > 0$,

$$d_i \succ_q d_j \quad \Longleftrightarrow \quad \widehat{p}_{d_i}(q) - \widehat{p}_{d_j}(q) \geq \delta. \tag{5}$$

Both documents belong to $D_k(q)$. Retained preferences form a triplet set $\mathcal{R}$ for noisy pairwise ranking supervision, while ties and smaller gaps remain unspecified. A threshold $\delta > 2\varepsilon_m(\eta)$ gives the simultaneous guarantee in Theorem 2. Our limited-budget setting uses $m = 4$ and $\delta = 0.25$, which retain pairs separated by at least one observation. This is an empirical filtering rule and does not provide that high-confidence guarantee.

The surrogate is a dual encoder with score $\widehat{s}_\theta(q, d) = \langle f^q_\theta(q), f^d_\theta(d) \rangle$. We preserve membership supervision from every observed original or perturbed query. Let $\mathcal{O}$ contain the cached records $(u, D_k(u))$, and let $\mathcal{N}(u)$ contain sampled documents available to the attacker that are absent from $D_k(u)$. With equal weight for each retrieved positive, the contrastive membership term is

$$\mathcal{L}_{\text{mem}} = - \sum_{(u, D_k(u)) \in \mathcal{O}} \sum_{d^+ \in D_k(u)} \log \frac{\exp \widehat{s}_\theta(u, d^+)}{\exp \widehat{s}_\theta(u, d^+) + \sum_{d^- \in \mathcal{N}(u)} \exp \widehat{s}_\theta(u, d^-)}. \tag{6}$$

We keep the membership objective and negative-sampling protocol fixed across the equal-budget comparisons. The joint training objective is

$$\begin{aligned} \mathcal{L}_{\text{sur}} &= \mathcal{L}_{\text{mem}} + \alpha \mathcal{L}_{\text{rank}}, \\ \mathcal{L}_{\text{rank}} &= \sum_{(q, d^+, d^-) \in \mathcal{R}} \left[m_r - \widehat{s}_\theta(q, d^+) + \widehat{s}_\theta(q, d^-) \right]_+ . \end{aligned} \tag{7}$$

Here $[x]_+ = \max(0, x)$, $m_r > 0$ is the ranking margin, and $\alpha > 0$ controls the pairwise term. Membership supervision teaches which documents enter the retrieved set. The pairwise term additionally supervises their relative ordering without requiring retention probabilities to match the target score scale.

3.4 Stability-Aware Trigger Optimization

Let d_{inj} be an injected document. For a target query q with cached probing observations, we select the b least stable documents in $D_k(q) \setminus \{d_{\text{inj}}\}$ as the estimated boundary competitor set $A(q)$. Stability ties are resolved by increasing surrogate score. These documents are candidates for lying near the retrieval cutoff under the preceding analysis. Their exact boundary positions remain unknown.

Using the cached query variations, we optimize a short query-side trigger T by minimizing

$$\mathcal{L}_{\text{trig}}(T \mid q) = \frac{1}{m} \sum_{t=1}^{m} \sum_{d_a \in A(q)} \left[m_t - \widehat{s}_\theta(q^{(t)} \oplus T, d_{\text{inj}}) + \widehat{s}_\theta(q^{(t)} \oplus T, d_a) \right]_+ . \tag{8}$$

The operator $\oplus$ appends the trigger and $m_t > 0$ is the trigger margin. Beam search adds or replaces trigger tokens using only surrogate scores. The objective promotes the injected document above estimated boundary competitors across local query variations. It provides a surrogate optimization target and does not guarantee top-k entry under the target retriever.

At attack execution, the submitted query is $q \oplus T$. A retrieved injected document enters the evidence supplied to the LLM, and attack success requires the generated response to adopt the attacker-specified content. Surrogate training and trigger search use no further target-retriever calls once their required observations are cached. Probing a new target query to construct $A(q)$ incurs additional calls. Appendix B.9 specifies the probing protocol and call accounting.

4 Experiments

4.1 Experimental Setup

Models and datasets. We evaluate nine RAG configurations combining Contriever Izacard et al. (2021), Co-Condenser Gao & Callan (2022), and ANCE Xiong et al. (2020) with Llama-3-8B-Instruct Grattafiori et al. (2024), Vicuna-7B-v1.5 Chiang et al. (2023), and Mistral-7B-Instruct-v0.2 Jiang et al. (2023). Retrieval uses inner-product similarity with $k = 10$. The attacker observes retrieved document identities and contents but not their scores or ordering. Target parameters and gradients remain inaccessible. We use Natural Questions Kwiatkowski et al. (2019) and TriviaQA Joshi et al. (2017) with a shared Wikipedia snapshot. For each dataset, a fixed split assigns 3,000 queries to surrogate construction and 2,000 disjoint queries to attack evaluation. The surrogate is frozen before attack evaluation.

Table 1: Surrogate fidelity under identical target observations at $B = 3$, comprising one original query and two perturbations. EW, FW, and RW denote Eq-Weight, Freq-Weight, and Rand-Weight.

Retriever	Supervision	Inter@10 ↑	τ_b@10 ↑	DCG@10 Gap ↓
Contriever	EW	0.604 (0.007)	0.154 (0.010)	0.808 (0.014)
	FW	0.610 (0.006)	0.210 (0.006)	0.740 (0.004)
	RW	0.594 (0.003)	0.135 (0.006)	0.835 (0.007)
	RV-KCA	**0.633 (0.007)**	**0.287 (0.011)**	**0.651 (0.014)**
Co-Condenser	EW	0.596 (0.011)	0.162 (0.012)	0.804 (0.018)
	FW	0.584 (0.010)	0.199 (0.018)	0.761 (0.023)
	RW	0.591 (0.015)	0.149 (0.017)	0.822 (0.026)
	RV-KCA	**0.618 (0.012)**	**0.290 (0.010)**	**0.653 (0.015)**
ANCE	EW	0.536 (0.014)	0.059 (0.022)	0.959 (0.033)
	FW	0.547 (0.012)	0.114 (0.015)	0.880 (0.020)
	RW	0.539 (0.015)	0.053 (0.020)	0.965 (0.031)
	RV-KCA	**0.575 (0.012)**	**0.204 (0.014)**	**0.760 (0.020)**

Baselines and attack setting. We compare RV-KCA with PAT Liu et al. (2022), FlippedRAG Chen et al. (2025), and GARAG Cho et al. (2024) under common black-box access and document-injection constraints. FlippedRAG is adapted from opinion manipulation to fact substitution. Each attack instance uses one 100–150-token injected document containing an attacker-specified false statement. All methods receive the same initial document and target statement. RV-KCA appends a query-side trigger of at most 16 wordpiece tokens.

Metrics and protocol. Attack Success Rate (ASR) is the fraction of responses that adopt the target statement, as judged by an independent Llama-3.1-8B-Instruct evaluator Grattafiori et al. (2024). Surrogate fidelity is measured by Inter@10 for set overlap, Kendall's τ_b@10 for ordering agreement, and DCG@10 Gap for top-weighted ranking discrepancy. Higher values are better for the first two metrics, while a lower gap is better. We average results over five paired runs with fixed data splits. One black-box call submits one query and returns an unordered top-10 set. Calls/query measures surrogate-construction calls, including original queries and their perturbations. Additional probing for attack-query boundary selection and final attack executions are accounted for separately. Surrogate training and trigger search require no further target calls once their required observations are cached. Appendix C provides implementation and evaluation details.

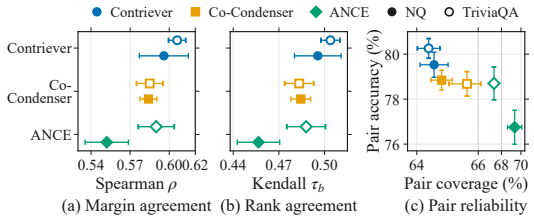


Figure 3: Evaluation of the stability signal on Natural Questions and TriviaQA, with pair accuracy reported alongside coverage.

4.2 Main Experimental Results

4.2.1 Retrieval Stability as a Ranking Signal

As shown in Figure 4, stronger probing lowers candidate acceptance and increases intent changes among accepted perturbations on both datasets. At the default $\lambda = 0.20$, over 95% of audited accepted perturbations preserve the original information need, supporting their use to study retrieval stability around the same question. Preservation declines more sharply at $\lambda = 0.30$, where document turnover may partly reflect changes in the question itself. This highlights why perturbation strength must be controlled when interpreting retention differences as ranking evidence. Lower acceptance also increases local resampling effort, while rejected candidates consume no target calls.

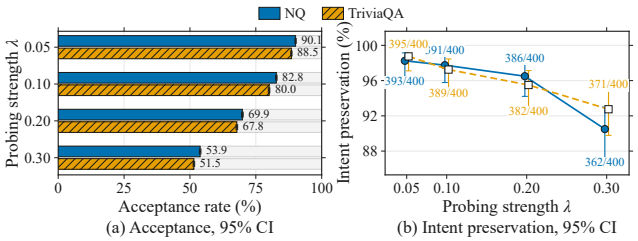


Figure 4: Candidate acceptance (a) and human-judged intent preservation among accepted perturbations (b). Each audit uses 400 samples per dataset and strength, and error bars show 95% confidence intervals.

Figure 3 shows that retrieval stability carries useful ranking information across both datasets and all three retrievers. Documents retained more often under query perturbations tend to have larger margins above the retrieval cutoff and rank nearer the top of the target list. The selected pairwise

Table 2: ASR (%) across nine RAG configurations. PAT-Q and FlippedRAG-Q are query-side adaptations, and Δ denotes RV-KCA's paired gain over FlippedRAG-Q in percentage points.

Generator	Retriever	PAT-Q	FlippedRAG-Q	GARAG	RV-KCA	Δ
Llama-3-8B-Instruct	Contriever	33.38 (0.38)	42.78 (1.47)	24.77 (0.99)	**50.27 (0.76)**	+7.5
	Co-Condenser	28.51 (1.04)	35.74 (0.66)	22.89 (1.34)	**43.17 (1.19)**	+7.4
	ANCE	31.38 (1.24)	39.16 (0.87)	24.92 (1.06)	**43.93 (1.79)**	+4.8
Vicuna-7B-v1.5	Contriever	42.02 (0.17)	50.80 (1.25)	31.76 (1.06)	**56.33 (0.99)**	+5.5
	Co-Condenser	34.89 (1.20)	43.23 (0.80)	29.04 (1.12)	**50.86 (1.21)**	+7.6
	ANCE	39.49 (1.45)	45.69 (1.58)	30.45 (2.54)	**51.43 (1.19)**	+5.7
Mistral-7B-Instruct-v0.2	Contriever	30.19 (0.67)	39.02 (0.57)	24.64 (0.94)	**46.77 (0.78)**	+7.8
	Co-Condenser	25.91 (0.51)	**33.96 (1.08)**	26.92 (0.75)	33.27 (1.43)	−0.7
	ANCE	29.70 (1.26)	34.76 (1.28)	27.76 (2.23)	**38.59 (0.66)**	+3.8
Macro average		32.83 (0.35)	40.57 (0.59)	27.02 (1.09)	**46.07 (0.81)**	+5.5

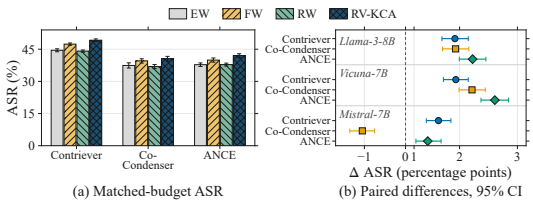
preferences agree with the target ordering for roughly four out of five pairs while covering about two-thirds of the available pairs. This combination suggests that stability provides useful supervision over a substantial fraction of document pairs. These findings support RV-KCA's use of stability differences to infer pairwise preferences from unordered retrieval outputs, while the remaining errors show that this signal is imperfect.

4.2.2 Surrogate Fidelity under Identical Observations

RV-KCA achieves the best surrogate fidelity across all three retrievers under identical target observations, as shown in Table 1. Frequency weighting improves ranking agreement over equal and random weighting, suggesting that document retention carries useful information about target preferences. RV-KCA further improves both ranking metrics by converting stability differences into explicit pairwise supervision. Its higher top-10 overlap shows that these gains also preserve and improve agreement on which documents are retrieved. With observations and training settings matched, these results support our claim that retrieval stability provides useful ranking supervision from unordered outputs, even with only two perturbations per query. Appendix D.3 reports the complete $B \in \{3, 4, 5\}$ comparison.

4.2.3 Transfer to End-to-End Attacks

Table 2 indicates that RV-KCA achieves the highest ASR in eight of the nine configurations, with a macro-average gain of 5.5 percentage points over the strongest baseline, FlippedRAG-Q. The gains across different retrievers and generators support the effectiveness of RV-KCA beyond a single model pairing. Combined with the surrogate fidelity results, they support our claim that ranking information inferred from retrieval stability can help influence the generator's final answer. The exception is Co-Condenser paired with Mistral, where RV-KCA slightly trails FlippedRAG-Q. This highlights a limit to the downstream benefit, since successful corruption also requires the generator to adopt the injected content. Dataset-specific results and the retrieval-to-generation breakdown appear in Table 3.



(a) Matched-budget ASR (b) Paired differences, 95% CI

Figure 5: Attack performance at $B = 3$ under matched observations, varying only the frozen surrogate. (a) reports ASR by retriever, averaged over datasets and generators with run standard deviations, while (b) shows RV-KCA's paired gains over FW for each configuration with 95% query-bootstrap intervals.

As illustrated in Figure 5, RV-KCA achieves the highest average ASR for each retriever under matched observations. It outperforms frequency weighting in eight of the nine configurations, with the corresponding paired confidence intervals entirely above zero. The reversal for Co-Condenser paired with Mistral shows that improved surrogate fidelity does not guarantee higher attack success for every model pairing. Since all variants share target observations and trigger-search settings, changing only the frozen surrogate isolates the contribution of surrogate supervision. The overall

Table 3: Dataset-level decomposition of attack success by retrieval outcome.

Dataset	Method	Hit@10	$\mathrm{ASR}_{\mathrm{hit}}$	$\mathrm{ASR}_{\mathrm{miss}}$	ASR	Δ_{clean}
NQ	PAT-Q	52.07 (0.67)	65.77 (0.89)	0.76 (0.08)	34.63 (0.55)	33.92 (0.55)
	FlippedRAG-Q	61.50 (1.09)	68.37 (0.37)	**0.84 (0.09)**	42.40 (0.69)	41.68 (0.69)
	GARAG	39.94 (1.77)	**70.61 (1.33)**	0.74 (0.07)	28.61 (1.40)	27.90 (1.40)
	RV-KCA	**70.20 (1.17)**	67.76 (0.52)	0.78 (0.10)	**47.86 (0.72)**	**47.15 (0.72)**
TriviaQA	PAT-Q	47.77 (0.32)	64.13 (0.43)	0.71 (0.10)	31.02 (0.34)	30.13 (0.34)
	FlippedRAG-Q	57.55 (0.95)	66.75 (0.39)	0.70 (0.06)	38.74 (0.55)	37.85 (0.55)
	GARAG	36.23 (1.32)	**68.96 (0.85)**	**0.74 (0.08)**	25.42 (0.85)	24.52 (0.85)
	RV-KCA	**66.39 (1.28)**	66.26 (0.63)	0.67 (0.16)	**44.27 (0.95)**	**43.38 (0.95)**

gains therefore support our claim that stability-derived pairwise preferences provide useful attack guidance under a fixed interaction budget.

Table 3 demonstrates that RV-KCA's advantage comes mainly from retrieving the injected document more often. Its conditional ASR after successful retrieval is slightly below FlippedRAG-Q on both datasets, yet its higher Hit@10 accompanies the strongest overall attack performance. GARAG illustrates this distinction by achieving the highest conditional ASR but the lowest retrieval success. The low ASR on retrieval misses and the substantial gains over clean inputs further link answer corruption to exposure to the injected content. These findings support RV-KCA's central mechanism of using stability-derived ranking information to make controlled documents more likely to reach the generator.

4.2.4 Contribution of Boundary-Aware Trigger Search

Surrogate supervision and boundary selection make complementary contributions to attack effectiveness, as illustrated in Table 4. With the search objective fixed, the stability-trained surrogate improves both Hit@10 and ASR, supporting the value of recovered ranking information for trigger optimization. Enabling boundary selection produces further gains with the same surrogate, suggesting that targeting competitors near the retrieval cutoff helps injected documents enter the retrieved set. The accompanying ASR gains show that both changes also improve success in corrupting the final answer. Since probing records and search budgets are matched, these comparisons support the contributions of both components under the same resource constraints.

Table 4: Attack-side component analysis at $B = 3$.

Variant	Boundary selection	Hit@10 ↑	ASR ↑
Membership-trained surrogate	Off	55.76 (0.96)	37.72 (0.60)
Stability-trained surrogate	Off	61.00 (0.98)	41.21 (0.73)
Full RV-KCA	On	**65.05 (0.98)**	**43.93 (0.73)**

4.2.5 Efficiency of Surrogate Construction

In Table 5, RV-KCA reaches the same surrogate fidelity targets with fewer construction calls on average across all three retrievers. The largest reduction occurs for ANCE, where FlippedRAG-Q requires the most observations, suggesting that stability-derived supervision is particularly useful when the target retriever is harder to imitate. This complements the fixed-budget results by showing that the ranking signal extracted from retrieval stability also reduces the cost of reaching a given fidelity level. These savings support our claim of efficient surrogate construction and apply specifically to construction calls, since additional probing and final attack executions are excluded.

Table 5: Construction calls C_{sur} to reach Inter@10 $\geq$ 0.60 and DCG@10 Gap $\leq$ 1.50 on the fixed diagnostic subset, with $N = 3{,}000$ construction queries and additional probing and final executions excluded.

Retriever	FlippedRAG-Q ↓	RV-KCA ↓	Reduction (%) ↑
Contriever	7,020 (1,211)	**5,760 (311)**	15.8 (16.1)
Co-Condenser	9,060 (1,957)	**7,020 (1,252)**	21.4 (10.9)
ANCE	16,710 (1,506)	**11,220 (1,708)**	**32.9 (6.7)**

Figure 6 indicates that RV-KCA achieves higher ASR than FlippedRAG-Q at matched construction budgets across all three retrievers. It also reaches comparable attack success with fewer construction calls, extending the efficiency advantage to downstream attacks. Gains diminish near the default budget, although very early checkpoints remain less effective. This shows that reducing construction

cost can still sacrifice attack strength, even when a surrogate already meets a fidelity threshold. By evaluating each checkpoint's own attack performance, these curves complement the fixed-fidelity comparison and support the practical value of learning useful surrogates from fewer target observations.

4.3 Response Log-Perplexity Analysis

We use response log-perplexity to characterize changes in generated responses. The same Llama-3.1-8B checkpoint and tokenizer score both clean and attacked responses. Clean responses are generated for the same queries without document injection or a query-side trigger. The metric definition is provided in Appendix D.8.

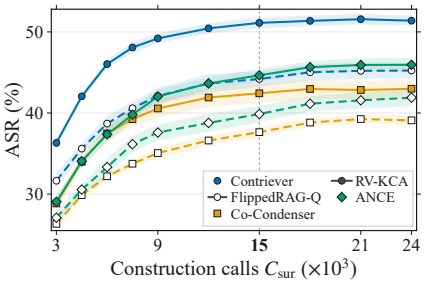


Figure 6: ASR achieved using each surrogate checkpoint under fixed instance-probing and trigger-search settings. The dashed line marks the default budget of 15,000 calls.

Figure 7 demonstrates that RV-KCA's response log-PPL distributions largely overlap the clean distributions on both datasets, with a modest shift toward higher values. Its curves closely follow FlippedRAG-Q, while GARAG exhibits a more pronounced upper tail.

To quantify these differences, we compute the empirical 1-Wasserstein distance between clean and attacked response log-PPL distributions within each configuration, then macro-average these distances within each run. We report the mean and standard deviation across five runs. Two-sided 95% confidence intervals use Student's t distribution with four degrees of freedom. Each interval is the mean plus or minus $2.776\,\mathrm{SD}/\sqrt{5}$, where SD denotes the standard deviation across runs.

Table 6 shows that RV-KCA and FlippedRAG-Q have similar distances from the clean distributions, both smaller than those of PAT-Q and GARAG. Together with RV-KCA's higher ASR, these results show that more frequent factual corruption can coexist with relatively small changes in response log-PPL. This finding highlights the need to assess factual corruption separately from response likelihood. Small distributional differences do not establish semantic preservation or resistance to detection.

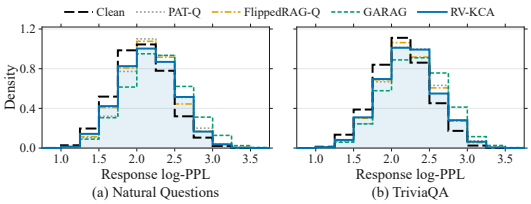


Figure 7: Response log-PPL distributions for paired clean and attacked responses on Natural Questions and TriviaQA, evaluated using the same fixed scoring model.

Table 6: Macro-averaged 1-Wasserstein distance W between attacked and paired clean response log-PPL distributions.

Method	Mean W	Run SD	95% CI
PAT-Q	0.113	0.008	[0.103, 0.123]
FlippedRAG-Q	<u>0.085</u>	0.013	[0.069, 0.101]
GARAG	0.183	0.013	[0.167, 0.199]
RV-KCA	**0.083**	0.008	[0.073, 0.093]

5 Conclusion and Future Work

This work shows that unordered top-k outputs can expose ranking information through retrieval stability in black-box RAG. Our analysis establishes a monotonic relationship between retention probability and retrieval margin under a shared local noise model and bounds estimation error under independent sampling. Building on this connection, RV-KCA converts stability under semantics-preserving query perturbations into pairwise preferences that complement membership supervision for margin-aware surrogate training. The surrogate then guides query-side trigger optimization to promote injected documents over estimated boundary competitors across query variations. Experiments on two benchmarks show higher surrogate fidelity and overall attack success across multiple retrievers and LLMs. RV-KCA also reaches fixed fidelity targets with fewer target-retriever calls for surrogate construction. These findings demonstrate that hiding similarity scores and explicit ranks does not eliminate exploitable ranking information from retrieval outputs. Future work should develop retrieval defenses that reduce this behavioral leakage while preserving retrieval utility.

6 AI USE STATEMENT

AI tools were used for language editing and text polishing during manuscript preparation. The study also includes experiments evaluating large language models, as described in the experimental setup. The authors take full responsibility for the final manuscript, experimental methodology, and reported results.

REFERENCES

Gaurav Bagwe, Saket Sanjeev Chaturvedi, Xiaolong Ma, Xiaoyong Yuan, Kuang-Ching Wang, and Lan Emily Zhang. Your rag is unfair: Exposing fairness vulnerabilities in retrieval-augmented generation via backdoor attacks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15930–15948, 2025.

Harsh Chaudhari, Giorgio Severi, John Abascal, Matthew Jagielski, Christopher A Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. Phantom: General trigger attacks on retrieval augmented language generation. *arXiv preprint arXiv:2405.20485*, 2024.

Zhuo Chen, Yuyang Gong, Jiawei Liu, Miaokun Chen, Haotan Liu, Qikai Cheng, Fan Zhang, Wei Lu, and Xiaozhong Liu. Flippedrag: Black-box opinion manipulation adversarial attacks to retrieval-augmented generation models. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 4109–4123, 2025.

Pengzhou Cheng, Yidong Ding, Tianjie Ju, Zongru Wu, Wei Du, Ping Yi, Zhuosheng Zhang, and Gongshen Liu. Trojanrag: Retrieval-augmented generation can be backdoor driver in large language models. *arXiv preprint arXiv:2405.13401*, 2024.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

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, 2024.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 31–36, 2018.

Luyu Gao and Jamie Callan. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2843–2853, 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. In *Neural Information Processing Systems*. Curran Associates, 2024.

Saidakhror Gulyamov, Said Gulyamov, Andrey Rodionov, Rustam Khursanov, Kambariddin Mekhmonov, Djakhongir Babaev, and Akmaljon Rakhimjonov. Prompt injection attacks in large language models and ai agent systems: A comprehensive review of vulnerabilities, attack vectors, and defense mechanisms. *Information*, 17(1):54, 2026.

Minseon Gwak, Min Su Kim, Jong Pil Yun, and PooGyeon Park. Robust and explainable fault diagnosis with power-perturbation-based decision boundary analysis of deep learning models. *IEEE Transactions on Industrial Informatics*, 19(5):6982–6992, 2022.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Yang Jiao, Xiaodong Wang, and Kai Yang. Pr-attack: Coordinated prompt-rag attacks on retrieval-augmented generation in large language models via bilevel optimization. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 656–667, 2025.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Chunyang Li, Junwei Zhang, Anda Cheng, Zhuo Ma, Xinghua Li, and Jianfeng Ma. Cpa-rag: Covert poisoning attacks on retrieval-augmented generation in large language models. *arXiv preprint arXiv:2505.19864*, 2025a.

Hao Li, Jiajun He, Guangshuo Wang, Dengguo Feng, Zheng Li, and Min Zhang. Budgetleak: Membership inference attacks on rag systems via the generation budget side channel. *arXiv preprint arXiv:2511.12043*, 2025b.

Xun Liang, Simin Niu, Zhiyu Li, Sensen Zhang, Hanyu Wang, Feiyu Xiong, Zhaoxin Fan, Bo Tang, Jihao Zhao, Jiawei Yang, et al. Saferag: benchmarking security in retrieval-augmented generation of large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4609–4631, 2025.

Jiawei Liu, Yangyang Kang, Di Tang, Kaisong Song, Changlong Sun, Xiaofeng Wang, Wei Lu, and Xiaozhong Liu. Order-disorder: Imitation adversarial attacks for black-box neural ranking models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2025–2039, 2022.

Yash More, Prakhar Ganesh, and Golnoosh Farnadi. Towards more realistic extraction attacks: An adversarial perspective. *Transactions of the Association for Computational Linguistics*, 13:1832–1849, 2025.

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

Agada Joseph Oche, Ademola Glory Folashade, Tirthankar Ghosal, and Arpan Biswas. A systematic review of key retrieval-augmented generation (rag) systems: Progress, gaps, and future directions. *arXiv preprint arXiv:2507.18910*, 2025.

Kaiyi Pang, Tao Qi, Chuhan Wu, Minhao Bai, Minghu Jiang, and Yongfeng Huang. Modelshield: adaptive and robust watermark against model extraction attack. *IEEE Transactions on Information Forensics and Security*, 20:1767–1782, 2025.

Congzheng Song, Alexander M Rush, and Vitaly Shmatikov. Adversarial semantic collisions. *arXiv preprint arXiv:2011.04743*, 2020.

Feiyang Wang, Xingquan Zuo, Hai Huang, and Gang Chen. Adba: approximation decision boundary approach for black-box adversarial attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 7628–7636, 2025a.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, et al. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17716–17736, 2024.

Yuhao Wang, Wenjie Qu, Shengfang Zhai, Yanze Jiang, Zichen Liu, Yue Liu, Yinpeng Dong, and Jiaheng Zhang. Silent leaks: Implicit knowledge extraction attack on rag systems through benign queries. *arXiv preprint arXiv:2505.15420*, 2025b.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten De Rijke, Yixing Fan, and Xueqi Cheng. Prada: Practical black-box adversarial attacks against neural ranking models. *ACM Transactions on Information Systems*, 41(4):1–27, 2023.

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.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

Jiaqi Xue, Mengxin Zheng, Yebowen Hu, Fei Liu, Xun Chen, and Qian Lou. Badrag: Identifying vulnerabilities in retrieval augmented generation of large language models. *arXiv preprint arXiv:2406.00083*, 2024.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37:121156–121184, 2024.

Quan Zhang, Binqi Zeng, Chijin Zhou, Gwihwan Go, Heyuan Shi, and Yu Jiang. Human-imperceptible retrieval poisoning attacks in llm-powered applications. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, pp. 502–506, 2024.

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, 2025.

A Additional Related Work

A.1 Adversarial Attacks on Retrieval Models

Adversarial attacks on neural retrieval systems differ in the information available to the attacker. Gradient-based text attacks provide a foundation for studying these vulnerabilities. HotFlip (Ebrahimi et al., 2018) uses gradients to identify token substitutions that alter a text classifier's predictions. Collision (Song et al., 2020) combines gradient-based search with language-model constraints to construct adversarial sequences for semantic matching. Such sequences can promote irrelevant documents in retrieval rankings. These studies demonstrate the vulnerability of text models to adversarial perturbations, but direct access to target gradients limits their applicability to deployed retrieval services.

Black-box ranking attacks exploit observable outputs to construct transferable perturbations. PAT (Liu et al., 2022) learns an imitation ranker and uses a pairwise objective to optimize triggers that promote a target document. PRADA (Wu et al., 2023) learns a surrogate from pseudo relevance feedback and uses its gradients to guide word substitutions. Its feedback includes rank positions within a partial retrieved list. These methods show that retrieval outputs can support surrogate learning and transferable attacks. The supervision available to the attacker, however, depends on the information exposed by the retrieval interface.

RV-KCA considers an unordered top-k interface that hides both similarity scores and rank positions. A single response supplies membership labels without specifying the relative order of retrieved documents. RV-KCA extracts additional evidence by measuring document retention under semantics-preserving query perturbations. Our analysis establishes that retention probability is non-decreasing

in the original retrieval margin under a shared local noise model. Empirical stability differences provide noisy ordering evidence under this condition. RV-KCA filters these differences into pairwise preferences that complement membership supervision for a margin-aware surrogate. This signal provides partial ordering information without reconstructing absolute target scores. The perturbations serve as diagnostic probes, while a separately optimized query-side trigger is used for attack execution.

A.2 Security and Reliability of Retrieval-Augmented Generation

Retrieval-Augmented Generation improves factual grounding by conditioning generation on external evidence (Oche et al., 2025; Wang et al., 2024). This dependence also creates a security risk because corrupted evidence can redirect generation without modifying the generator itself (Xue et al., 2024; Zou et al., 2025).

Prior work has identified several ways to manipulate the evidence supplied to a RAG system. PoisonedRAG (Zou et al., 2025) injects crafted documents to induce attacker-specified answers. BadRAG (Xue et al., 2024) and Phantom (Chaudhari et al., 2024) investigate attacks that associate triggers with malicious retrieval outcomes and downstream behavior. GARAG (Cho et al., 2024) studies low-level document perturbations that disrupt RAG responses. Human-imperceptible retrieval poisoning further shows that documents can mislead generation while appearing benign to readers (Zhang et al., 2024). Together, these studies establish that retrieved content is an effective attack surface even when the generator itself remains unchanged.

Some attacks explicitly coordinate retrieval manipulation with downstream generation. PR-Attack (Jiao et al., 2025) formulates the optimization of poisoned texts and prompt triggers as a bilevel problem. FlippedRAG (Chen et al., 2025) learns a surrogate retriever from black-box observations and uses it to optimize poisoning triggers for opinion manipulation. FlippedRAG is particularly relevant to RV-KCA because it connects retriever approximation to end-to-end attack effectiveness. RV-KCA investigates how changes in unordered top-k membership can supply additional ranking supervision for this approximation.

Security benchmarks and defense methods address complementary aspects of this threat. SafeRAG (Liang et al., 2025) provides a benchmark for evaluating RAG security under different forms of corrupted retrieved context. RobustRAG (Xiang et al., 2024) mitigates retrieval corruption through an isolate-then-aggregate strategy. These studies evaluate or limit the effects of corrupted evidence. RV-KCA complements them by examining ranking information exposed through retrieval behavior before generation.

RV-KCA contributes both a source of ranking evidence and a way to use that evidence for surrogate training. It measures document persistence under controlled query variations and converts stability differences into pairwise preferences. These preferences augment membership supervision, and the resulting surrogate guides query-side trigger optimization against estimated boundary competitors. Repeated top-k observations can therefore support stronger knowledge corruption attacks even when similarity scores and explicit ranks remain hidden.

B Additional Details of RV-KCA

This appendix provides the proofs and implementation details for Section 3. It first states the conditions that connect retrieval stability to hidden document ordering. The remaining sections specify how the observed top-k sets support surrogate training and trigger optimization. Target scores and margins are used only for analysis. All attack observations come from submitted queries and the corresponding unordered retrieved sets.

B.1 Theoretical Analysis

The analysis conditions on a fixed query q and a fixed corpus $\mathcal{C}$. The target retrieval cutoff and the margin of $d \in D_k(q)$ are

$$\tau_k(q) = \mathrm{kth}\{s_\star(q,d) \mid d \in \mathcal{C}\}\,, \tag{9}$$

$$\gamma_d(q) = s_\star(q,d) - \tau_k(q), \tag{10}$$

where kth denotes the k-th largest score. The deterministic retention result permits score ties. The ordering results additionally assume no ties between documents at the cutoff, as stated in Theorem 1. A fixed tie-breaking rule makes membership well defined in the implementation, but does not remove this additional assumption from the ordering analysis.

B.1.1 A Sufficient Condition for Retention

We first formalize the intuition that a sufficiently large retrieval margin protects a document against bounded score changes. Let ϕ be the target query encoder. Suppose that, for every document $d \in \mathcal{C}$ and every accepted perturbation q' in the local neighborhood,

$$|s_\star(q', d) - s_\star(q, d)| \le L \left\|\phi(q') - \phi(q)\right\|. \tag{11}$$

This bound must hold over the entire corpus because documents outside the original top-k set can change the perturbed cutoff. The target encoder and the constant L need not be available to the attacker.

Lemma 1 (Stability of the retrieval cutoff). *If every target score changes by at most ϵ, then*

$$|\tau_k(q') - \tau_k(q)| \le \epsilon. \tag{12}$$

Proof. At least k documents have original scores at least $\tau_k(q)$. Each has perturbed score at least $\tau_k(q) - \epsilon$, so $\tau_k(q') \ge \tau_k(q) - \epsilon$. At most $k - 1$ documents have original scores strictly above $\tau_k(q)$. A document with perturbed score strictly above $\tau_k(q) + \epsilon$ must belong to this group. Thus $\tau_k(q') \le \tau_k(q) + \epsilon$.

Proposition 1 (Margin-based retention). *Under Equation 11, suppose $\|\phi(q') - \phi(q)\| \le r$. For every $d \in D_k(q)$,*

$$\gamma_d(q) > 2Lr \quad \Longrightarrow \quad d \in D_k(q'). \tag{13}$$

Proof. Set $\epsilon = Lr$. Equation 11 and Lemma 1 give

$$\begin{aligned} s_\star(q', d) - \tau_k(q') &\ge s_\star(q, d) - \epsilon - \tau_k(q) - \epsilon \\ &= \gamma_d(q) - 2Lr. \end{aligned} \tag{14}$$

The assumed strict inequality makes the perturbed score exceed the cutoff, which ensures membership even if other documents tie.

The proposition is a sufficient condition. A document with a smaller margin can also remain retrieved when its score changes favorably relative to its competitors. Therefore, this deterministic result alone does not justify inferring document ordering from observed retention. The shared local noise model supplies the additional condition used in Theorem 1. Semantic filtering also does not establish a numerical bound on the inaccessible target embedding displacement.

B.1.2 Proof and Scope of the Ordering Result

For a document $d \in D_k(q)$, the noise in Equation 2 can be written explicitly as

$$Z_d(q') = \big[s_\star(q', d) - s_\star(q, d)\big] - \big[\tau_k(q') - \tau_k(q)\big]. \tag{15}$$

The cutoff changes with the full corpus. Sharing the distribution of the document-score change alone would not suffice for Theorem 1. The assumption concerns the complete relative change in Equation 15.

Proof of Theorem 1. In the absence of score ties between documents at the cutoff,

$$d \in D_k(q') \quad \Longleftrightarrow \quad s_\star(q', d) - \tau_k(q') \ge 0. \tag{16}$$

The k-th document itself has zero perturbed margin and is included. Substituting Equation 2 gives

$$p_d(q) = \Pr(Z_d(q') \geq -\gamma_d(q)) = \Pr(Z \geq -\gamma_d(q)) \,. \tag{17}$$

If $\gamma_{d_i}(q) \geq \gamma_{d_j}(q)$, then the event $\{Z \geq -\gamma_{d_j}(q)\}$ is contained in $\{Z \geq -\gamma_{d_i}(q)\}$. Hence $p_{d_i}(q) \geq p_{d_j}(q)$. Conversely, if $p_{d_i}(q) > p_{d_j}(q)$, the inequality $\gamma_{d_i}(q) \leq \gamma_{d_j}(q)$ would contradict this monotonicity. Thus $\gamma_{d_i}(q) > \gamma_{d_j}(q)$. Subtracting the common original cutoff preserves score ordering.

The proof uses only a common marginal distribution for the compared documents. It does not require independent score changes across documents or Gaussian noise. Discrete noise is permitted. If F is the common cumulative distribution function, the exact identity is $p_d(q) = 1 - F(-\gamma_d(q)^-)$, where $F(x^-) = \Pr(Z < x)$. The left limit accounts for any mass at the retention threshold.

Monotonicity need not be strict. For $\gamma_{d_i}(q) > \gamma_{d_j}(q)$, the difference in retention probabilities equals

$$p_{d_i}(q) - p_{d_j}(q) = \Pr\big(-\gamma_{d_i}(q) \leq Z < -\gamma_{d_j}(q)\big) \,. \tag{18}$$

A strict probability difference therefore requires positive probability in this interval. Equal retention probabilities do not establish equal target scores. This is why RV-KCA leaves tied empirical stabilities unordered.

Bounded score changes alone do not imply the ordering result. Consider $k = 2$ and three documents with original scores $(1.02, 1.01, 1.00)$. Their original top-2 margins are 0.01 and 0. Suppose a perturbation changes the scores to $(0.99, 1.01, 1.00)$ with probability 0.6 and leaves them unchanged otherwise. The first document has retention probability 0.4, while the second is always retained. Every score change is bounded by 0.03, yet stability reverses their original ordering. The example establishes the need for an additional condition linking document sensitivities. It does not claim that the shared local noise model is the only sufficient condition.

B.1.3 Finite-Sample Estimation and Pairwise Reliability

Proof of Theorem 2. Condition on the original query and its retrieved set. For each fixed $d \in D_k(q)$, define $X_{t,d} = \mathbf{1}\{d \in D_k(q^{(t)})\}$. Independent perturbations make $X_{1,d}, \ldots, X_{m,d}$ independent Bernoulli variables with mean $p_d(q)$. Thus $\mathbb{E}[\widehat{p}_d(q)] = p_d(q)$. Hoeffding's inequality Hoeffding (1963) gives, for every $e > 0$,

$$\Pr(|\widehat{p}_d(q) - p_d(q)| \geq e) \leq 2\exp(-2me^2). \tag{19}$$

A union bound over the k original documents yields

$$\Pr\left(\max_{d \in D_k(q)} |\widehat{p}_d(q) - p_d(q)| \geq e\right) \leq 2k\exp(-2me^2).$$

Taking $e = \varepsilon_m(\eta)$ gives Equation 4. On its simultaneous event,

$$\big|(\widehat{p}_{d_i} - \widehat{p}_{d_j}) - (p_{d_i} - p_{d_j})\big| \leq 2\varepsilon_m(\eta). \tag{20}$$

The dependence on q is omitted here for readability. An empirical difference exceeding this bound implies a positive population difference. Theorem 1 then gives the target-score ordering under its assumptions.

Membership events for different documents within the same probe can be dependent. The proof requires independence across probes for a fixed document, and the union bound requires no independence across documents. Because Equation 20 holds simultaneously for all original documents, it also covers pairs selected after observing their empirical stabilities.

Probe requirements. For a fixed threshold δ, the sufficient condition $\delta > 2\varepsilon_m(\eta)$ is equivalent to

$$m > \frac{2}{\delta^2} \log \frac{2k}{\eta}. \tag{21}$$

This is a sufficient sample requirement for simultaneous retention-order certification within one query. It is not a bound on the number of calls needed to reach a particular surrogate fidelity or

attack success rate. For N queried neighborhoods, a union bound gives a simultaneous statement by replacing $\log(2k/\eta)$ with $\log(2Nk/\eta)$ when k and m are shared.

For the reported setting $k = 10$ and $m = 4$, taking $\eta = 0.05$ gives $2\varepsilon_m(\eta) \approx 1.73$. Since an empirical probability difference is at most one, the distribution-free bound certifies no pair in this setting. The empirical threshold $\delta = 0.25$ therefore controls the resolution of the supplied supervision. Its practical reliability must be assessed empirically. The conservativeness of the bound does not imply that the observed pairs are uninformative.

A bound for a fixed pair. Let $\Delta = p_{d_i}(q) - p_{d_j}(q)$ and $\widehat{\Delta} = \widehat{p}_{d_i}(q) - \widehat{p}_{d_j}(q)$. Set $Y_t = X_{t,d_i} - X_{t,d_j} \in [-1, 1]$. Applying the one-sided Hoeffding inequality to these independent variables gives

$$\Pr(\widehat{\Delta} - \Delta \geq a) \leq \exp(-ma^2/2), \qquad \Pr(\Delta - \widehat{\Delta} \geq a) \leq \exp(-ma^2/2) \tag{22}$$

for $a > 0$. If $\Delta \leq 0$, the probability of incorrectly retaining the direction $d_i \succ_q d_j$ is at most $\exp(-m\delta^2/2)$. If $\Delta > \delta$, the probability of failing to retain this direction is at most $\exp[-m(\Delta - \delta)^2/2]$. These bounds express the trade-off between filtering incorrect directions and retaining weak but correct preferences. A bound for one fixed pair alone does not certify all data-selected pairs.

B.1.4 Fixed Allocation Across Perturbation Families

The implementation allocates a fixed number of probes to each perturbation family. Let family h supply m_h independent accepted perturbations, with $\sum_h m_h = m$. Define its retention probability as $p_{d,h}(q)$ and set

$$p_d(q) = \sum_h \frac{m_h}{m} p_{d,h}(q). \tag{23}$$

The pooled estimator in Equation 1 remains unbiased for this allocation-weighted quantity. The indicators are independent across probes but need not have the same mean. Hoeffding's inequality for independent bounded variables gives the same bound in Equation 19, so the simultaneous and pairwise results remain valid.

For the ordering result, the shared local noise assumption must hold under the allocation-weighted perturbation distribution. A sufficient condition is a common noise distribution across documents within each family, with the same family weights for all compared documents. The distributions may differ between families.

Family allocation and acceptance rules are fixed before observing target retrieval outcomes. Independent candidate generation followed by a fixed semantic filter gives independent accepted draws within each family. Reusing dependent candidates or choosing later probes from earlier target responses does not automatically satisfy the preceding concentration argument. Such protocols require a separate analysis. The fixed four-family design uses one accepted perturbation per family and defines $p_d(q)$ as their equally weighted population retention probability.

B.1.5 Sensitivity to Document-Dependent Noise

The shared local noise model is an idealization. We next quantify how a bounded departure from that model changes the ordering condition. This extension is an analytical statement and introduces no additional observations into RV-KCA.

For the fixed query, let $H_d(z) = \Pr(Z_d \geq z)$. Suppose there is a reference tail function $H(z) = \Pr(Z \geq z)$ such that

$$\sup_z |H_d(z) - H(z)| \leq \rho \tag{24}$$

for every compared document, where $\rho \geq 0$. Retain the no-ties assumption from Theorem 1.

Proposition 2 (Ordering under bounded noise differences). *Under Equation 24, $p_{d_i}(q) - p_{d_j}(q) > 2\rho$ implies $s_\star(q, d_i) > s_\star(q, d_j)$. On the event in Theorem 2, the empirical condition*

$$\widehat{p}_{d_i}(q) - \widehat{p}_{d_j}(q) > 2\varepsilon_m(\eta) + 2\rho \tag{25}$$

therefore certifies the same target-score direction.

Proof. Write $g(\gamma) = H(-\gamma)$, which is non-decreasing. Equation 24 implies $|p_d(q) - g(\gamma_d(q))| \le \rho$. If $\gamma_{d_i}(q) \le \gamma_{d_j}(q)$, then

$$p_{d_i}(q) - p_{d_j}(q) \le g(\gamma_{d_i}(q)) - g(\gamma_{d_j}(q)) + 2\rho \le 2\rho. \tag{26}$$

The first claim follows by contradiction. Equation 20 lowers an empirical difference by at most $2\varepsilon_m(\eta)$, giving the second claim.

This proposition separates sampling error from differences in document sensitivity. Additional independent probes reduce the former. They do not eliminate the latter. The quantity ρ is not identified by top-k membership alone, so this extension is not an operational certificate without an independently justified bound on ρ.

B.2 Threat Model and Scope

Observable interface. The attacker can submit text queries and observe the identities of the returned documents. The contents of observed documents are available for surrogate training. The interface reveals the unordered set $D_k(q)$ but hides target similarity scores and explicit ranks. Target parameters and gradients remain inaccessible. The attacker also has no access to target embeddings or index internals and does not control retriever training or indexing hyperparameters.

Attacker-controlled inputs. The attacker can inject a bounded number of controlled documents into the external corpus and submit a target query augmented with a query-side trigger. This is a query-triggered knowledge corruption setting. Applying the trigger requires control over, or influence on, the submitted query. It does not assume access to an immutable third-party query.

Each injected document contains the attacker-specified target statement and follows the common 100–150-token template. The target system indexes it through the permitted document-insertion interface. Benign documents cannot be modified or deleted. Compared methods use the same injection budget and document template. They also share the target statements and queries, with an identical indexing protocol. The corpus is held fixed during each original-query and perturbed-query probing block. Cached target observations are reused only for the same corpus state and probing configuration.

Objective and success criterion. For a target query q and injected document d_{inj}, the retrieval objective is to promote d_{inj} into $D_k(q \oplus T)$. Retrieval makes the injected content available to the generator. An end-to-end attack succeeds only when the generated response adopts or supports the attacker-specified content under the experimental evaluation protocol. Retrieval success alone does not establish generation success.

B.3 Semantics-Preserving Perturbation Neighborhood

RV-KCA generates local query variations from four perturbation families. A common semantic filter is applied before any candidate is submitted to the target. The filter restricts changes in retrieval intent, while the perturbation strength controls the extent of the proposed surface change.

Lexical substitution. For $q = (w_1, \ldots, w_n)$, let $\mathcal{I}(q)$ contain eligible token positions. Named entities and numbers are protected. Domain-specific expressions and syntactic heads that determine the requested information are also excluded. Replacing an eligible token with a context-appropriate synonym gives

$$q' = (w_1, \ldots, w_i', \ldots, w_n), \qquad i \in \mathcal{I}(q). \tag{27}$$

Structural rephrasing. Let $\mathcal{T}$ contain transformations that preserve the information need, such as clause reordering and equivalent question forms. A structural candidate is

$$q' = g(q), \qquad g \in \mathcal{T}. \tag{28}$$

Active–passive alternation is permitted when it preserves the roles of the entities in the query.

Contextual expansion. A neutral fragment ζ is appended to form

$$q' = q \oplus \zeta. \tag{29}$$

The fragment may specify a neutral response format or background context. It must not introduce a new entity or factual claim, or change an answer constraint.

Anchor-term suppression. Let $\mathcal{H}(q)$ contain nonessential retrieval-sensitive tokens identified from term rarity and syntactic role. Protected terms are excluded from this set. Removing a subset $\mathcal{H}_s \subseteq \mathcal{H}(q)$ gives

$$q' = q \setminus \mathcal{H}_s. \tag{30}$$

Here the set difference denotes token deletion from the query sequence. This operation probes sensitivity to lexical anchors while retaining terms needed to express the information request.

Semantic-preservation filter. Each candidate is compared with its source query using cosine similarity under `text-embedding-3-large`. A candidate is accepted only if

$$\mathrm{sim}_{\mathrm{sem}}(q, q') \geq 0.80. \tag{31}$$

The protocol additionally rejects changes to the requested entity or relation. It also rejects changes to temporal qualifiers and answer constraints. Passing the embedding threshold alone does not guarantee semantic equivalence. The audit found that more than 95% of sampled accepted perturbations preserved the information need. The audit protocol and sample size are reported in the experimental appendix. These checks support the use of local query variations but do not establish the target-encoder smoothness or shared-noise assumptions.

B.4 Perturbation Strength and Query Budget

The probing strength $\lambda \in [0, 1]$ controls candidate generation before semantic filtering. The default setting is $\lambda = 0.20$.

- **Lexical substitution.** Replace at most $\lceil \lambda n_{\mathrm{rep}} \rceil$ eligible tokens, where $n_{\mathrm{rep}} = |\mathcal{I}(q)|$.
- **Structural rephrasing.** Permit stronger transformations as λ increases, subject to the same semantic filter.
- **Contextual expansion.** Append at most $\lceil 20\lambda \rceil$ neutral tokens.
- **Anchor-term suppression.** Remove at most $\lceil \lambda |\mathcal{H}(q)| \rceil$ nonessential anchor tokens.

The semantic similarity threshold remains fixed at 0.80 for every strength.

For $\lambda > 0$, the per-query target-retriever budget B includes the original query. We distinguish the population perturbation distribution $\mathcal{P}_\lambda(q)$ from the finite list of submitted perturbations,

$$\mathcal{V}_{\lambda,B}(q) = (q^{(1)}, \ldots, q^{(m_q)}), \qquad 1 \leq m_q \leq B - 1. \tag{32}$$

The notation $m_q = |\mathcal{V}_{\lambda,B}(q)|$ counts list entries. Repeated draws remain separate observations if submitted separately. For one fixed query, m_q is the quantity denoted by m in the main text.

Family allocation is fixed independently of target responses. Accepted candidates are generated independently within each family by applying a fixed filter to independent proposals. The standard $B = 5$ design submits one accepted perturbation from each family. For $B \in \{3, 4\}$, a predetermined subset of these family slots is used, with allocations balanced across construction queries. If more probes are required, additional accepted candidates are generated under the same fixed allocation rule. For a given budget, the family weights are part of the distribution $\mathcal{P}_\lambda(q)$. Section B.1.4 states the corresponding design-averaged interpretation of $p_d(q)$.

Equal-budget comparisons use the same original and perturbed top-k observations. They differ in how these observations supervise the surrogate. The setting $\lambda = 0$ is the membership-only baseline. It uses the original top-k set and defines no perturbation-based stability estimate. The stability and boundary-selection steps below require $m_q \geq 1$. When probing strengths are compared, all target calls issued for each configuration are included in its reported cost. Results from different perturbation distributions are not pooled into one stability estimate without specifying their weights.

B.5 Stability Annotation and Pair Construction

For each construction query q, RV-KCA records $D_k(q)$ and the response to every entry in $\mathcal{V}_{\lambda,B}(q)$. For $d \in D_k(q)$, the empirical stability is

$$\widehat{p}_d(q) = \frac{1}{m_q} \sum_{t=1}^{m_q} \mathbf{1}\Big\{ d \in D_k(q^{(t)}) \Big\}. \tag{33}$$

This is the same statistic as Equation 1. It measures retention within the local probing design and is not calibrated to the target similarity scale.

For $d_i, d_j \in D_k(q)$, the retained preference is

$$d_i \succ_q d_j \quad \Longleftrightarrow \quad \widehat{p}_{d_i}(q) - \widehat{p}_{d_j}(q) \geq \delta. \tag{34}$$

Each retained preference contributes (q, d_i, d_j) to $\mathcal{R}$. Ties and gaps below δ contribute no ordering constraint. A document can participate in several pairs. If a query yields no pair, its observations still contribute membership supervision.

The threshold is fixed before evaluation. Under the shared local noise model, $\delta > 2\varepsilon_{m_q}(\eta)$ provides the simultaneous ordering guarantee in Theorem 2. The default $m_q = 4$ and $\delta = 0.25$ retain pairs separated by at least one observation. This practical rule does not satisfy the high-confidence condition discussed in Section B.1.3. Raising δ reduces the set of retained pairs and tightens the fixed-pair error bound, but does not by itself validate the shared-noise assumption.

Cached observations. We distinguish two views of the same probing data. The collection $\mathcal{O}$ contains all observed query-response pairs $(u, D_k(u))$ used by the membership objective. The grouped cache $\mathcal{K}(q)$ stores the original response and the perturbation list for each probed query. It also stores every perturbed response and the resulting values $\widehat{p}_d(q)$. This grouping introduces no additional target information. Keeping the full perturbed responses is necessary because documents newly retrieved under $q^{(t)}$ also provide membership supervision for that query.

B.6 Margin-Aware Surrogate Training

The surrogate uses query encoder f_θ^q and document encoder f_θ^d, with score

$$\widehat{s}_\theta(q, d) = \left\langle f_\theta^q(q), f_\theta^d(d) \right\rangle. \tag{35}$$

For the triplet collection $\mathcal{R}$, the pairwise loss and total objective are

$$\mathcal{L}_{\text{rank}} = \sum_{(q, d^+, d^-) \in \mathcal{R}} \left[m_r - \widehat{s}_\theta(q, d^+) + \widehat{s}_\theta(q, d^-) \right]_+, \tag{36}$$

$$\mathcal{L}_{\text{sur}} = \mathcal{L}_{\text{mem}} + \alpha \mathcal{L}_{\text{rank}}. \tag{37}$$

Here $\mathcal{L}_{\text{mem}}$ is exactly the membership objective in Equation 6, applied to the full observation collection $\mathcal{O}$. For each observed query u, negatives are sampled from documents available to the attacker and excluded from $D_k(u)$. Absence from this set is a membership label and does not imply semantic irrelevance.

The membership objective and negative-sampling protocol are held fixed across equal-budget variants. Surrogate initialization and the training schedule are also shared. Membership supervision learns which documents enter the observed top-k sets. The pairwise term adds preferences between documents in the same original retrieved set. The margin $m_r > 0$ requires a nonzero surrogate score separation for each retained preference. These surrogate margins have no prescribed numerical equivalence to target-score margins.

The reported configuration initializes a `bert-base-uncased` dual encoder and uses $m_r = 0.20$. Once construction is complete, the surrogate is frozen for trigger search. Additional target-query probing for boundary selection does not update its parameters.

B.7 Boundary-Competitor Selection

For a target query with a compatible cached probing record, RV-KCA selects estimated boundary competitors from the original retrieved set,

$$A(q) = \mathrm{Bottom}_b\left(\left\{\left(d, \widehat{p}_d(q)\right) \mid d \in D_k(q) \setminus \{d_{\mathrm{inj}}\}\right\}\right). \tag{38}$$

The operator returns up to b document identities with the smallest empirical stability. Ties are resolved by increasing $\widehat{s}_\theta(q, d)$, followed by a fixed document-identifier rule. The size parameter b is distinct from the target-call budget B. The default uses $k = 10$ and $b = 3$.

Excluding d_{inj} prevents comparison of the injected document with itself, which would otherwise contribute the constant positive loss m_t. Trigger search requires a nonempty competitor set. Under the ordering assumptions, low retention probabilities suggest small original retrieval margins. Finite-sample error and document-dependent noise limit this interpretation, so $A(q)$ is an estimated competitor set. It need not equal the documents nearest the true cutoff.

The construction requires the original and perturbed target responses for the same query q. A construction-query cache supplies these quantities only when the query and probing configuration match and the corpus is unchanged. For a new target query, the required observations are collected before trigger search and counted as additional probing. A competitor set estimated solely from surrogate retrieval is a different construction and is not substituted for $A(q)$ in the stated method.

B.8 Stability-Aware Trigger Search

Fix an attack instance (q, d_{inj}) and its cached probing record. For a local variation q' and a competitor $d_a \in A(q)$, define

$$\Delta_T(q', d_a) = m_t - \widehat{s}_\theta(q' \oplus T, d_{\mathrm{inj}}) + \widehat{s}_\theta(q' \oplus T, d_a). \tag{39}$$

The per-query trigger objective is

$$\mathcal{L}_{\mathrm{trig}}(T \mid q) = \frac{1}{m_q} \sum_{t=1}^{m_q} \sum_{d_a \in A(q)} \left[\Delta_T(q^{(t)}, d_a)\right]_+. \tag{40}$$

This is Equation 8 with $m = m_q$. The objective is averaged over the m_q accepted perturbations.

A query-specific attack optimizes one trigger for one target query. For a shared trigger targeting one fixed injected document over a set $\mathcal{Q}_{\mathrm{trig}}$, the objective is $\sum_{q \in \mathcal{Q}_{\mathrm{trig}}} \mathcal{L}_{\mathrm{trig}}(T \mid q)$. Each query uses its own probing record and competitor set. The query-specific case has $|\mathcal{Q}_{\mathrm{trig}}| = 1$. Algorithm 3 gives this case, which matches the main-text objective.

Search constraints and stopping. Let K be the beam width, $L_{\max}$ the token-length limit, and J the maximum number of search updates. These are separate constraints. An update proposes token additions or replacements and discards candidates longer than $L_{\max}$. Replacements remain possible after the length limit is reached. Candidates are scored with the frozen surrogate, and the K lowest-loss candidates are retained. The previous beam is included in the candidate pool so that the best available trigger is not lost.

Search stops after J updates or when the best trigger satisfies $\Delta_T(q^{(t)}, d_a) \le 0$ for every cached variation and selected competitor. This condition is equivalent to zero loss in Equation 40 and is checked on the full finite objective. It guarantees the specified surrogate margins on the cached comparisons only. It does not certify target top-k entry or generation success.

The reported settings use $K = 8$ and $m_t = 0.10$. The search evaluates 32 proposed token candidates per update, permits at most $L_{\max} = 16$ trigger tokens, and uses at most $J = 40$ updates. Candidate proposal and scoring use only the surrogate. After search, the final query is submitted once to the target RAG system. A trigger optimized from a record for q is evaluated on that same query unless a separate transfer protocol is explicitly specified.

B.9 Complete Attack Procedure and Call Accounting

Algorithm 1 summarizes surrogate construction followed by attack evaluation with a frozen surrogate. Algorithm 2 trains the surrogate once on the fixed benign corpus without injected attack

documents, using the construction-query set $\mathcal{Q}_{\mathrm{sur}}$. For each evaluation instance, Algorithm 3 uses the matching target-query record for trigger search and performs the final attack execution. If a compatible record is unavailable after document injection, it is acquired separately while the surrogate remains frozen. This record supports boundary selection and trigger search and is not added to the surrogate-training data. After each instance, the injected document is removed to restore the benign index.

Algorithm 1 RV-KCA under an unordered top-k interface

Require: Target retrieval interface $R_\star$, target RAG system, construction queries $\mathcal{Q}_{\mathrm{sur}}$, evaluation instances (q, d_{inj})
Require: Retrieval depth k, per-query budget $B \geq 2$, strength $\lambda > 0$, threshold $\delta > 0$, training and search settings
Ensure: Frozen surrogate parameters θ, triggers and responses for all evaluation instances

Start from the fixed benign Wikipedia index without injected attack documents.
Run Algorithm 2 once to obtain θ, observations $\mathcal{O}$, and grouped cache $\mathcal{K}$.
Freeze θ.
for each evaluation instance (q, d_{inj}) **do**
  Inject d_{inj} through the permitted interface and hold the indexed corpus fixed for this instance.
  if a compatible complete record $\mathcal{K}(q)$ is unavailable **then**
    Query $R_\star(q)$ and record $D_k(q)$.
    Generate the fixed-allocation list $\mathcal{V}_{\lambda,B}(q)$ using Sections B.3 and B.4.
    Query $R_\star(q^{(t)})$ for every list entry and retain all returned sets.
    Compute $\widehat{p}_d(q)$ for $d \in D_k(q)$ using Equation 33.
    Store these observations in $\mathcal{K}(q)$ without updating θ or the training collection $\mathcal{O}$.
    Add $1 + m_q$ calls to the additional target-query probing cost.
  end if
  Run Algorithm 3 with the record $\mathcal{K}(q)$ to obtain $T^\star$ and y.
  Record $(q, T^\star, y)$.
  Remove d_{inj} and restore the benign index.
end for
return θ and the recorded triggers and responses

Construction and additional probing costs. Let $N = |\mathcal{Q}_{\mathrm{sur}}|$. Under one fixed probing configuration, constructing the surrogate observations costs

$$C_{\mathrm{sur}} = \sum_{q \in \mathcal{Q}_{\mathrm{sur}}} (1 + m_q) \leq NB. \tag{41}$$

Let $\mathcal{Q}_{\mathrm{extra}}$ contain target queries that require a new complete probing record after construction. For the same fixed corpus and configuration, the full observation cost is

$$\begin{aligned} C_{\mathrm{extra}} &= \sum_{q \in \mathcal{Q}_{\mathrm{extra}}} (1 + m_q), \\ C_{\mathrm{probe}} &= C_{\mathrm{sur}} + C_{\mathrm{extra}} \leq \big(N + |\mathcal{Q}_{\mathrm{extra}}|\big) B. \end{aligned} \tag{42}$$

A compatible complete record can be reused without a new target call. If a change in the corpus or probing configuration requires another record, the corresponding newly issued calls are counted again. Equation 42 applies to one fixed configuration, and costs are summed across configurations when new observations are collected.

Every submitted original or perturbed query counts as one target-retriever call. Rejected candidates that are never submitted use no target-retriever calls. Local filtering and surrogate computation are separate from this target-call metric. Training and trigger search use no further target calls once their required observations are cached. The final submission in Algorithm 3 uses one target-system call per attack instance and is reported separately from probing.

Equal-budget comparisons must match C_{sur} and also account for C_{extra} when attack-query probing is used. A previously unqueried test query cannot simultaneously use the target-derived $A(q)$ in

Algorithm 2 Black-box stability probing and surrogate training

Require: Target retrieval interface $R_\star$, fixed benign Wikipedia corpus without injected attack documents, construction queries $\mathcal{Q}_{\mathrm{sur}}$, depth k, budget $B \geq 2$, strength $\lambda > 0$, threshold $\delta > 0$
Ensure: Surrogate parameters θ, all training observations $\mathcal{O}$, grouped cache $\mathcal{K}$

Initialize $\mathcal{O}$ and $\mathcal{R}$ as empty collections and $\mathcal{K}$ as an empty cache.
for each $q \in \mathcal{Q}_{\mathrm{sur}}$ **do**
  Query $R_\star(q)$ once and record $D_k(q)$.
  Append $(q, D_k(q))$ to $\mathcal{O}$.
  Generate $\mathcal{V}_{\lambda,B}(q)$ with the fixed family allocation and semantic filter.
  for each entry $q^{(t)}$ in $\mathcal{V}_{\lambda,B}(q)$ **do**
    Query $R_\star(q^{(t)})$ once and record $D_k(q^{(t)})$.
    Append $(q^{(t)}, D_k(q^{(t)}))$ to $\mathcal{O}$.
  end for
  Compute $\widehat{p}_d(q)$ for each $d \in D_k(q)$ using Equation 33.
  for each ordered pair of distinct documents $(d_i, d_j) \in D_k(q) \times D_k(q)$ **do**
    if $\widehat{p}_{d_i}(q) - \widehat{p}_{d_j}(q) \geq \delta$ **then**
      Append (q, d_i, d_j) to $\mathcal{R}$.
    end if
  end for
  Store $D_k(q)$ and $\mathcal{V}_{\lambda,B}(q)$ in $\mathcal{K}(q)$, together with every perturbed response and all computed stabilities.
end for
Train the dual encoder on $\mathcal{O}$ and $\mathcal{R}$ by minimizing $\mathcal{L}_{\mathrm{mem}} + \alpha\mathcal{L}_{\mathrm{rank}}$.
return $\theta, \mathcal{O}, \mathcal{K}$

Algorithm 3 Stability-aware trigger search and attack execution

Require: Target RAG system, frozen surrogate $\widehat{s}_\theta$, target query q, its complete record $\mathcal{K}(q)$, injected document d_{inj}
Require: Competitor limit b, trigger margin m_t, beam width K, token limit $L_{\max}$, update limit J, initial trigger T_{init} with $|T_{\mathrm{init}}| \leq L_{\max}$
Ensure: Trigger $T^\star$ and final target response y

Construct the nonempty set $A(q)$ using Equation 38, excluding d_{inj}.
Initialize the beam $\mathcal{B}_0 \leftarrow \{T_{\mathrm{init}}\}$ and evaluate its loss using Equation 40.
Set $T^\star \leftarrow T_{\mathrm{init}}$.
for $j = 1$ to J **do**
  if $\mathcal{L}_{\mathrm{trig}}(T^\star \mid q) = 0$ **then**
    break
  end if
  Propose permitted token additions and replacements for candidates in $\mathcal{B}_{j-1}$.
  Form candidate pool $\mathcal{E}_j$ from these proposals and $\mathcal{B}_{j-1}$, retaining only triggers of length at most $L_{\max}$.
  for each $T \in \mathcal{E}_j$ **do**
    Evaluate $\mathcal{L}_{\mathrm{trig}}(T \mid q)$ on the cached variations and $A(q)$ using only $\widehat{s}_\theta$.
  end for
  Let $\mathcal{B}_j$ contain up to K lowest-loss candidates in $\mathcal{E}_j$.
  Set $T^\star \leftarrow \arg\min_{T \in \mathcal{B}_j} \mathcal{L}_{\mathrm{trig}}(T \mid q)$, with a fixed rule for ties.
end for
Submit $q \oplus T^\star$ to the target RAG system once and record its response y.
return $T^\star, y$

Equation 38 and incur zero additional target access. The distinction concerns how observations are acquired and counted. It does not change the assumption that all target scores and explicit ranks remain hidden.

C Detailed Experimental Setup

This section specifies the target systems and the shared experimental protocol. It also details surrogate construction and attack execution. The attack interface exposes unordered top-10 results. Target scores and ordered rankings retained for evaluation are not provided to the attack methods.

C.1 Models and Retrieval Systems

Target configurations. We combine Contriever Izacard et al. (2021), Co-Condenser Gao & Callan (2022), and ANCE Xiong et al. (2020) with Llama-3-8B-Instruct Grattafiori et al. (2024), Vicuna-7B-v1.5 Chiang et al. (2023), and Mistral-7B-Instruct-v0.2 Jiang et al. (2023). Each of the nine retriever–generator configurations is evaluated separately. Generators use greedy decoding with a maximum of 128 new tokens and their model-specific chat templates. The target retriever and generator remain inaccessible to the attacker.

Corpus and retrieval. We use the December 20, 2018 English Wikipedia snapshot distributed with the DPR resources. It contains 21,015,324 non-overlapping passages of at most 100 words. All retrievers use this passage collection, with a separate index for each retriever's embeddings. Benign document embeddings are precomputed and stored in 32-bit floating-point format. We perform exact inner-product search with FAISS and return the top-10 documents, following the scoring convention in prior RAG attack studies Chen et al. (2025); Zou et al. (2025). The target system encodes each injected document and adds it to the corresponding index through the permitted insertion interface.

Observable feedback. The returned document identifiers are randomly permuted before they are exposed to the attacker. Their associated text is available for surrogate training. The interface therefore reveals top-10 membership without revealing similarity scores or rank positions. Target embeddings and index internals are also inaccessible. The evaluation process separately retains the original target ordering for fidelity measurement. It is not an input to probing, surrogate training, or trigger search.

C.2 Datasets and Data Splits

We use Natural Questions Kwiatkowski et al. (2019) and TriviaQA Joshi et al. (2017) as factual question-answering benchmarks. For each dataset, we sample 5,000 queries after removing questions with empty reference answers or no matching Wikipedia passage. A fixed split contains 3,000 construction queries and 2,000 attack-evaluation queries. The split is shared across methods and remains fixed across runs.

Only the construction split is used to fit the surrogate. A fixed 300-query subset of this split is used for hyperparameter selection. After selection, hyperparameters are frozen and the final surrogate is trained on the full construction split. The attack-evaluation queries are not used for surrogate fitting or hyperparameter selection.

For each evaluation query, RV-KCA optimizes a query-specific trigger with the frozen surrogate. Boundary selection requires the original and perturbed target responses for that same query under the applicable corpus state. If a compatible record is unavailable, these responses are collected before trigger search and counted as additional attack-query probing. They are used to select boundary competitors without updating the surrogate. This query-specific access is permitted by the threat model and is distinct from training on the evaluation split.

C.3 Attack Instances and Injection Protocol

Each attack instance pairs an evaluation query with an attacker-specified false statement and one controlled document. The document presents the statement as a factual claim, followed by two or three topically related supporting sentences. Its length is restricted to 100–150 tokens. All methods start from the same document text and target statement before applying their respective trigger or perturbation procedures. Benign documents cannot be modified or deleted.

For RV-KCA, the controlled document is indexed before the instance-specific probing record is acquired. The corpus is held fixed during the original-query and perturbed-query probing block and

throughout the subsequent trigger search. The final trigger is appended to the original evaluation query and submitted to the target RAG system. We restore the index after each instance so that documents injected for different queries cannot interact. A cached record is reused only when its query and probing configuration match and the corpus state is unchanged.

The main experiments limit RV-KCA's query-side trigger to 16 wordpiece tokens. Trigger-length sensitivity is evaluated separately. This setting assumes that the attacker can both inject the controlled document and modify the submitted query.

C.4 Perturbation and Surrogate Configuration

Perturbation generation. The default probing strength is $\lambda = 0.20$. We use lexical substitution and structural rephrasing, together with neutral contextual expansion and suppression of nonessential anchor terms. Named entities and numbers are protected from replacement or deletion. We also protect syntactic heads and expressions that determine the requested information. Anchor terms are selected from high inverse-document-frequency content words after excluding protected terms. Contextual expansions are drawn from a fixed collection of neutral fragments that introduce no new factual claims or answer constraints.

Each candidate must have cosine similarity of at least 0.80 to the original query under `text-embedding-3-large`. Candidates that change the requested entity or relation are rejected. The filter also rejects changes to temporal qualifiers or answer constraints. Rejected candidates are resampled before any target call is issued. In an audit of 800 accepted perturbations, 96% were judged to preserve the original information need. Agreement between the two annotators was Cohen's $\kappa = 0.84$. This is the semantic-preservation rate among accepted perturbations, not the acceptance rate of the automatic filter.

Probing and supervision. The default budget is $B = 5$ calls per construction query, comprising one original query and four accepted perturbations. One perturbation is sampled from each family, with independent candidate generation within the fixed allocation. For each document in the original top-10 set, empirical retrieval stability is the fraction of the four perturbed responses that retain that document. The original response is not included in this fraction.

We retain an ordered document pair when its stability difference is at least $\delta = 0.25$. With four perturbations, this corresponds to a difference of at least one retention observation. Ties and smaller differences provide no pairwise supervision. This threshold is an empirical resolution rule and does not satisfy the high-confidence condition in the finite-sample theorem.

The surrogate is a dual encoder initialized from `bert-base-uncased` and trained with the joint objective $\mathcal{L}_{\text{mem}} + \alpha\mathcal{L}_{\text{rank}}$ defined in the main paper. Membership supervision uses every cached original and perturbed query response, including documents retrieved only by a perturbation. For an observed query, negative documents are sampled from documents available to the attacker that are absent from its retrieved set. The pairwise term uses the retained preferences within each original top-10 set. Queries without retained pairs still contribute to the membership objective. Table 7 summarizes the reported training and search settings.

Boundary selection and trigger search. For an evaluation query with a compatible probing record, we select the three least stable documents in its original top-10 set after excluding the injected document. Stability ties are resolved by increasing surrogate score, followed by a fixed document-identifier rule. These documents form the estimated boundary competitor set. They are selected from target membership observations, not from a surrogate-only top-10 list.

The trigger objective averages over the cached accepted perturbations for that query. Beam search retains eight candidates and evaluates 32 proposed token candidates per update using only the frozen surrogate. Token additions and replacements respect the 16-token limit. The previous beam remains in the candidate pool. Search stops after 40 updates or when the injected document exceeds every selected competitor by $m_t = 0.10$ for every cached perturbation. This stopping condition concerns surrogate margins and does not certify retrieval by the target system. After the required probing record has been acquired, trigger search issues no further target-retriever calls.

Table 7: Reported RV-KCA implementation settings. Unless otherwise stated, these settings are shared across target retrievers.

Component	Setting
Target retrieval	Exact inner-product search, $k = 10$
Default probing	$\lambda = 0.20$, $B = 5$ calls per query
Semantic similarity threshold	0.80
Surrogate input limits	64 query / 256 document wordpiece tokens
Surrogate initialization	`bert-base-uncased` dual encoder
Surrogate representation	768-dimensional [CLS] embedding
Optimizer	AdamW, learning rate 2×10^{-5}
Training schedule	3 epochs, 10% linear warmup
Batch size	128 query–document pairs
Weight decay / gradient clipping	0.01 / 1.0
Pair threshold δ	0.25
Ranking margin m_r	0.20
Trigger margin m_t	0.10
Boundary competitor count	$b = 3$, excluding the injected document
Trigger search	Beam size 8, 32 token candidates per update
Maximum trigger length	16 wordpiece tokens
Maximum search updates	40
Injection budget	One 100–150-token document per instance
Independent runs	Five paired random seeds

C.5 Baseline Methods and Equal-Budget Controls

PAT. PAT Liu et al. (2022) optimizes a short adversarial trigger with a surrogate ranking model to promote a target document. We use its released optimization procedure subject to the common black-box interface and a 16-token trigger limit.

FlippedRAG. FlippedRAG Chen et al. (2025) approximates the target retriever from black-box retrieval observations and uses the surrogate to optimize adversarial triggers. We adapt its opinion-manipulation objective to fact substitution and use the same target statements as RV-KCA. Its implementation is evaluated under the common unordered top-10 interface and document-injection constraints.

Query-side adaptation. To match RV-KCA's query-triggered setting, we adapt PAT and FlippedRAG by moving their triggers from the document to the query. Each method optimizes a query-side trigger with its surrogate to promote a fixed injected document. Both adaptations use the same 16-token trigger limit and unordered top-10 interface as RV-KCA. The reported results refer to these query-side adaptations.

GARAG. GARAG Cho et al. (2024) uses genetic search to construct adversarial document perturbations that disrupt RAG. In this comparison, modifications are restricted to the controlled document under the shared document-length and injection constraints. Its generated responses are evaluated against the same attacker-specified statements.

All methods use the same query splits and corpus snapshot. They share the initial document text and target statements, with the same permitted black-box access. Target similarity scores and explicit ranks are unavailable to every method, as are target parameters and gradients. Hyperparameters are fixed before attack evaluation.

Equal-budget supervision controls. We compare Eq-Weight (EW), Freq-Weight (FW), and Rand-Weight (RW) with RV-KCA using identical original and perturbed retrieval observations. Eq-Weight assigns equal weight to retrieved positives and uses membership supervision. Freq-Weight uses empirical retention frequencies as document weights. Rand-Weight permutes these weights among documents in each original top-10 set. RV-KCA retains membership supervision and adds filtered pairwise preferences derived from stability differences.

The controls share the dual-encoder architecture and initialization. They also use the same training schedule and negative-sampling protocol. The perturbation queries and resulting target observations are identical across variants. This comparison tests how the observations are converted into supervision under a matched construction budget. Additional attack-query probing is accounted for separately when comparing downstream ASR.

C.6 Attack Success and Response Evaluation

Let $\mathcal{Q}_{\text{test}}$ denote the evaluation queries. For query q, let y_q be the response under the evaluated attack and let a_q^{adv} be the attacker-specified target statement. The evaluator returns $E(y_q, a_q^{\text{adv}}) \in \{0, 1\}$, and ASR is

$$\text{ASR} = \frac{1}{|\mathcal{Q}_{\text{test}}|} \sum_{q \in \mathcal{Q}_{\text{test}}} E(y_q, a_q^{\text{adv}}). \tag{43}$$

A positive label requires the response to accept or support the target statement. A response that merely mentions or quotes it is negative. Uncertainty and explicit rejection are also negative.

Following the use of independent response assessment in prior RAG attack evaluation Chen et al. (2025), we use Llama-3.1-8B-Instruct Grattafiori et al. (2024) as the evaluator. It receives the response and target statement and returns one binary label using greedy decoding. The evaluator is used only to measure ASR and supplies no feedback to the attack.

We audit 600 evaluator decisions sampled across the nine RAG configurations and four attack methods. Two annotators independently apply the same decision rule. Their agreement is 94.2%, with Cohen's $\kappa = 0.87$. The automatic evaluator agrees with the human reference labels on 92.8% of the audited responses.

Retrieval Success Rate is the fraction of evaluation queries for which the injected document enters the target top-10 set. Conditional ASR is computed over only those queries with successful retrieval. These metrics distinguish retrieval entry from adoption of the target statement after entry. Clean ASR uses the same queries and target statements without document injection or a trigger. We report attacked ASR and its absolute increase over clean ASR. When ASR is expressed as a percentage, this increase is reported in percentage points.

C.7 Surrogate Fidelity Metrics

For evaluation query q, let $\pi_{\text{tar}}^{10}(q)$ and $\pi_{\text{sur}}^{10}(q)$ be the ordered top-10 lists of the target and surrogate retrievers. Their corresponding document sets are $\mathcal{D}_{\text{tar}}^{10}(q)$ and $\mathcal{D}_{\text{sur}}^{10}(q)$. The two retrievers are compared on the same query and corpus state. The ordered target list is available only to the evaluation process.

Inter@10. Top-10 membership agreement is

$$\text{Inter@10}(q) = \frac{\left|\mathcal{D}_{\text{sur}}^{10}(q) \cap \mathcal{D}_{\text{tar}}^{10}(q)\right|}{10}. \tag{44}$$

Higher values indicate greater set overlap. This metric does not measure ordering.

Kendall's τ_b@10. We compute Kendall's τ_b on the union of the two retrieved sets. A document absent from one list is assigned rank 11 in that ranking, so missing documents are tied. The τ_b correction accounts for these ties. Higher values indicate stronger agreement between the two induced rankings. Because missing documents enter the calculation, the metric reflects membership differences as well as ordering.

DCG@10 Gap. We use target-rank gains to measure top-weighted ranking discrepancy. A document at target rank $r_{\text{tar}}(d)$ receives gain $g_q(d) = 1/\log_2(r_{\text{tar}}(d)+1)$. Documents outside the target top-10 receive zero gain. For an ordered list π, we define

$$\text{DCG@10}(\pi \mid q) = \sum_{i=1}^{10} \frac{g_q(\pi_i)}{\log_2(i+1)}. \tag{45}$$

The gap is

$$\mathrm{Gap}_{\mathrm{DCG@10}}(q) = \left|\mathrm{DCG@10}(\pi^{10}_{\mathrm{sur}}(q) \mid q) - \mathrm{DCG@10}(\pi^{10}_{\mathrm{tar}}(q) \mid q)\right|. \tag{46}$$

Lower values indicate closer agreement with the target ranking. The gain and position discounts emphasize discrepancies near the top of the list. This is a fidelity measure based on target ranks, not a measure based on external relevance judgments.

C.8 Black-Box Query Accounting

Surrogate construction. One black-box call submits one query to the target retriever and observes its unordered top-10 result. Let $\mathcal{Q}_{\mathrm{sur}}$ contain the N construction queries, and let m_q be the number of accepted perturbations actually submitted for query q. Under a per-query budget B, $m_q \le B-1$ and

$$C_{\mathrm{sur}} = \sum_{q \in \mathcal{Q}_{\mathrm{sur}}} (1 + m_q) \le NB, \qquad \text{Calls/query} = \frac{C_{\mathrm{sur}}}{N}. \tag{47}$$

The original query contributes one call. With $N = 3{,}000$ and exactly four perturbations per query, $B = 5$ gives $C_{\mathrm{sur}} = 15{,}000$ calls for one construction run on a dataset–retriever pair. This count covers surrogate construction only.

Additional probing and attack execution. Let $\mathcal{Q}_{\mathrm{extra}}$ contain attack queries that require a fresh complete probing record after construction. For one fixed probing configuration, the additional and combined probing costs are

$$C_{\mathrm{extra}} = \sum_{q \in \mathcal{Q}_{\mathrm{extra}}} (1 + m_q), \qquad C_{\mathrm{probe}} = C_{\mathrm{sur}} + C_{\mathrm{extra}}. \tag{48}$$

Each record must match the corpus state used for that attack instance. If a corpus change or a different probing configuration requires new observations, every newly issued call is counted again. Any repeated surrogate construction is also included in the construction cost.

Under the default protocol, we collect a fresh probing record for each of the 2,000 evaluation queries after document injection. Each record contains the response to the original query and four accepted perturbations, requiring 10,000 additional target-retriever calls per dataset-retriever pair. Together with the 15,000 construction calls, the total probing cost is 25,000 calls. Final attack execution adds 2,000 target-system calls per dataset-retriever-generator configuration, with one submission per evaluation instance. These execution calls are reported separately from probing. Probing records can be reused across generators when the query and retrieval conditions remain unchanged.

Candidate generation and semantic filtering do not query the target retriever. Surrogate training and trigger search also use no further target calls once their input observations are cached. Auxiliary-model computation is excluded from this target-retriever call metric. Rejected candidates incur no target calls because they are never submitted.

Equal-budget comparisons. For $B \in \{3, 4, 5\}$, the supervision controls receive identical original-query observations and exactly $B-1$ perturbed-query observations per construction query. When fewer than four probes are used, a fixed round-robin family allocation is shifted across queries to balance family coverage. The allocation is independent of target responses. At $\lambda = 0$, the membership-only configuration uses original-query observations and defines no perturbation-based stability estimate. Its cost is reported according to the calls actually issued.

A matched construction budget controls the amount of surrogate supervision. Comparisons of complete attack cost also include additional attack-query probing and final execution. Construction-only call reductions are therefore reported as surrogate-construction efficiency.

Fixed-fidelity efficiency. We begin with the 3,000 original-query observations and add accepted perturbation calls in a fixed shuffled order. The surrogate is updated after every 150 additional calls. This experiment varies the construction budget, so the default $B = 5$ is not a stopping cap. We score the resulting checkpoints on the construction selection subset and report the first checkpoint satisfying both $\mathrm{Inter@10} \ge 0.60$ and $\mathrm{Gap}_{\mathrm{DCG@10}} \le 1.5$. The thresholds and acquisition order are fixed before attack evaluation. The ordered target lists are used to measure checkpoint fidelity and

are not supplied to the surrogate or acquisition procedure. Reported costs in this experiment are C_{sur}.

C.9 Statistical Reporting

We use five paired runs, with fixed data splits and matched random seeds across methods. Metrics are first averaged over queries within each run and then summarized across runs. Run-to-run variation is measured by the standard deviation. Natural Questions and TriviaQA results are computed separately. Surrogate-fidelity aggregates give equal weight to the included dataset–retriever configurations. ASR aggregates additionally give equal weight to the included generators.

For paired method differences, we compute 95% confidence intervals using 10,000 bootstrap resamples of evaluation queries within each run, followed by aggregation across runs. Paired methods use the same resampled query indices. For significance tests across multiple retrievers, Holm's correction is applied within each metric and budget.

D Additional Experimental Results

The main text follows the evidence from the ranking signal to downstream use. The following analyses provide the complete budget comparisons and test the limits of the perturbation and search settings. Distributional diagnostics are reported separately from attack effectiveness.

D.1 Details of the Stability Diagnostic

For an original top-10 document d, the target margin is $\gamma_d(q) = s_\star(q, d) - \tau_{10}(q)$, where $\tau_{10}(q)$ is the tenth-largest target score. Within each query, we compute Spearman's correlation between empirical stability and the target margin, and Kendall's τ_b between the stability-induced order and the target order. Documents are ordered by decreasing empirical stability. Each correlation is averaged over queries for which it is defined. Queries with a constant stability vector are excluded from these correlation averages.

Retained-pair accuracy is the fraction of retained preferences consistent with the strict target-score order. Pairs with tied target scores are excluded from the accuracy calculation. Coverage is the number of retained unordered pairs divided by the number of available unordered pairs in the original top-10 sets. Pair accuracy is aggregated over retained pairs with unequal target scores, while the correlation summaries give equal weight to valid queries. Diagnostic target scores are never provided to an attack method.

Table 8: Validation of retrieval stability as a ranking signal.

Dataset	Retriever	Margin ρ	Rank τ_b	Pair accuracy (%)	Coverage (%)
NQ	Contriever	**0.596 (0.019)**	**0.496 (0.015)**	**79.53 (0.56)**	64.56 (0.46)
	Co-Condenser	0.584 (0.006)	0.484 (0.006)	78.84 (0.44)	64.81 (0.35)
	ANCE	0.552 (0.017)	0.457 (0.014)	76.75 (0.76)	**69.35 (0.73)**
TriviaQA	Contriever	**0.606 (0.007)**	**0.504 (0.006)**	**80.26 (0.43)**	64.39 (0.37)
	Co-Condenser	0.585 (0.010)	0.483 (0.010)	78.68 (0.54)	65.63 (0.58)
	ANCE	0.590 (0.014)	0.488 (0.013)	78.70 (0.73)	**67.35 (0.27)**

Table 8 shows that retrieval stability is positively associated with target margins and rankings across both datasets. This supports the premise that documents farther above the retrieval cutoff tend to survive query perturbations more often. Contriever yields the most accurate retained preferences, while ANCE retains more pairs but shows weaker agreement on NQ. This variation highlights why coverage must be assessed alongside accuracy when evaluating the supervision signal. Across all settings, retained preferences cover a substantial fraction of available pairs and mostly agree with the target ordering. These findings support RV-KCA's use of stability differences as informative but imperfect pairwise supervision.

Table 9 examines whether the association between retrieval stability and document identity contributes to surrogate fidelity. Shuffled-Pair performs similarly to EW, while RV-KCA achieves higher

Table 9: Stability-permutation control at $B = 3$.

Retriever	Supervision	Inter@10 ↑	τ_b@10 ↑	DCG@10 Gap ↓
Contriever	EW	0.604 (0.007)	0.154 (0.010)	0.808 (0.014)
	FW	0.610 (0.006)	0.210 (0.006)	0.740 (0.004)
	Shuffled-Pair	0.606 (0.009)	0.164 (0.014)	0.803 (0.019)
	RV-KCA	**0.633 (0.007)**	**0.287 (0.011)**	**0.651 (0.014)**
Co-Condenser	EW	0.596 (0.011)	0.162 (0.012)	0.804 (0.018)
	FW	0.584 (0.010)	0.199 (0.018)	0.761 (0.023)
	Shuffled-Pair	0.590 (0.013)	0.170 (0.017)	0.811 (0.022)
	RV-KCA	**0.618 (0.012)**	**0.290 (0.010)**	**0.653 (0.015)**
ANCE	EW	0.536 (0.014)	0.059 (0.022)	0.959 (0.033)
	FW	0.547 (0.012)	0.114 (0.015)	0.880 (0.020)
	Shuffled-Pair	0.540 (0.015)	0.054 (0.024)	0.949 (0.030)
	RV-KCA	**0.575 (0.012)**	**0.204 (0.014)**	**0.760 (0.020)**

Table 10: Candidate-generation settings for the probing-strength sweep. The semantic acceptance threshold is fixed at 0.80 for all nonzero strengths.

λ	Replacement fraction	Expansion limit	Anchor fraction	Semantic threshold
0.00	0	0 tokens	0	Not applied
0.05	0.05	1 token	0.05	0.80
0.10	0.10	2 tokens	0.10	0.80
0.20	0.20	4 tokens	0.20	0.80
0.30	0.30	6 tokens	0.30	0.80

fidelity across all three retrievers. RV-KCA consistently achieves higher surrogate fidelity while sharing the same pairwise objective and preference count with Shuffled-Pair. This comparison illustrates that the useful supervision comes from how stability estimates correspond to individual documents, supporting RV-KCA's use of this association to learn relative document ordering.

D.2 Perturbation Strength and Semantic Preservation

Probing strength controls candidate construction, while the acceptance rule stays fixed. Let n_{rep} be the number of eligible replacement positions and let $H(q)$ contain nonessential anchor terms. Lexical substitution changes at most $\lceil \lambda n_{\text{rep}} \rceil$ eligible tokens. Contextual expansion appends at most $\lceil 20\lambda \rceil$ neutral tokens, and anchor suppression deletes at most $\lceil \lambda |H(q)| \rceil$ eligible anchors. Structural rephrasing uses the configured strength-specific transformation set, specified as clause reordering and equivalent question forms, together with role-preserving active–passive alternation. At most $\lceil 10\lambda \rceil$ applicable transformations are composed per candidate.

Table 10 specifies how probing strength controls the query variations used to estimate retrieval stability. Larger λ values permit more extensive edits. Replacement and anchor-deletion counts are rounded up after applying the specified fractions to eligible tokens. Keeping the semantic threshold fixed allows us to examine how probing strength affects the ranking signal under a consistent acceptance rule. The $\lambda = 0$ setting provides a reference using only original-query membership supervision. This design tests whether controlled query

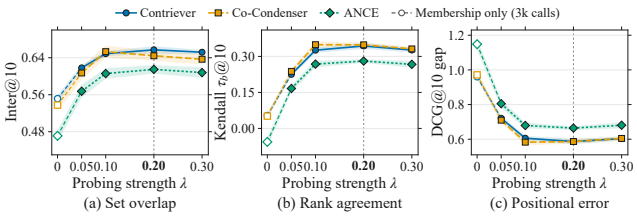


Figure 8: Surrogate fidelity as probing strength varies. Nonzero strengths use four accepted perturbations and the same semantic threshold. Each retriever is shown separately to expose metric-specific or retriever-specific trade-offs.

Table 11: Complete surrogate-fidelity comparison under identical observations at $B \in \{3, 4, 5\}$.

		Inter@10 ↑				τ_b@10 ↑				DCG@10 Gap ↓			
Model	B	EW	FW	RW	RV-KCA	EW	FW	RW	RV-KCA	EW	FW	RW	RV-KCA
Contriever	3	0.604 (0.007)	0.610 (0.006)	0.594 (0.003)	**0.633 (0.007)**	0.154 (0.010)	0.210 (0.006)	0.135 (0.006)	**0.287 (0.011)**	0.808 (0.014)	0.740 (0.004)	0.835 (0.007)	**0.651 (0.014)**
	4	0.612 (0.008)	0.622 (0.004)	0.598 (0.009)	**0.652 (0.007)**	0.173 (0.010)	0.242 (0.007)	0.137 (0.005)	**0.326 (0.010)**	0.782 (0.014)	0.701 (0.004)	0.830 (0.010)	**0.605 (0.012)**
	5	0.622 (0.008)	0.631 (0.007)	0.599 (0.007)	**0.657 (0.008)**	0.189 (0.008)	0.261 (0.005)	0.134 (0.002)	**0.344 (0.013)**	0.760 (0.011)	0.676 (0.006)	0.833 (0.004)	**0.586 (0.015)**
Co-Condenser	3	0.596 (0.011)	0.584 (0.010)	0.591 (0.015)	**0.618 (0.012)**	0.162 (0.012)	0.199 (0.018)	0.149 (0.017)	**0.290 (0.010)**	0.804 (0.018)	0.761 (0.023)	0.822 (0.026)	**0.653 (0.015)**
	4	0.605 (0.009)	0.603 (0.011)	0.592 (0.017)	**0.637 (0.010)**	0.183 (0.010)	0.235 (0.020)	0.147 (0.018)	**0.329 (0.008)**	0.776 (0.016)	0.714 (0.024)	0.823 (0.028)	**0.608 (0.011)**
	5	0.615 (0.011)	0.610 (0.010)	0.592 (0.014)	**0.644 (0.011)**	0.200 (0.012)	0.255 (0.020)	0.145 (0.015)	**0.348 (0.009)**	0.750 (0.017)	0.691 (0.024)	0.826 (0.023)	**0.587 (0.012)**
ANCE	3	0.536 (0.014)	0.547 (0.012)	0.539 (0.015)	**0.575 (0.012)**	0.059 (0.022)	0.114 (0.015)	0.053 (0.020)	**0.204 (0.014)**	0.959 (0.033)	0.880 (0.020)	0.965 (0.031)	**0.760 (0.020)**
	4	0.549 (0.014)	0.567 (0.012)	0.543 (0.015)	**0.599 (0.012)**	0.084 (0.021)	0.154 (0.014)	0.055 (0.019)	**0.250 (0.015)**	0.921 (0.031)	0.822 (0.019)	0.960 (0.030)	**0.702 (0.020)**
	5	0.564 (0.012)	0.581 (0.012)	0.544 (0.016)	**0.615 (0.009)**	0.106 (0.021)	0.180 (0.014)	0.054 (0.022)	**0.281 (0.008)**	0.886 (0.029)	0.787 (0.019)	0.961 (0.034)	**0.664 (0.011)**

variations supply useful ranking information and how sensitive that information is to perturbation strength.

As shown in Figure 8, surrogate fidelity improves with moderate probing, but further increases in strength offer little benefit. Contriever and ANCE perform best near $\lambda = 0.20$, while Co-Condenser favors 0.10. At 0.30, all three metrics deteriorate slightly despite an unchanged number of accepted probes. This pattern is consistent with the more frequent changes in query intent observed in the semantic audit. The results suggest that useful ranking supervision requires query variations that reveal retention differences while largely preserving the original information need. They support $\lambda = 0.20$ as a practical default, with the preferred strength varying across retrievers.

D.3 Complete Equal-Budget Fidelity Results

Table 11 shows that RV-KCA achieves the best results on all three fidelity metrics for every retriever and budget. Frequency weighting benefits from additional observations, yet its gap to RV-KCA persists, supporting the value of explicit pairwise supervision throughout the tested budget range. Random weighting shows little improvement as the budget grows, suggesting that useful supervision depends on preserving the link between each document and its observed stability. The simultaneous gains in retrieved-set overlap and ranking agreement indicate that RV-KCA better captures both which documents the target retrieves and how it orders them.

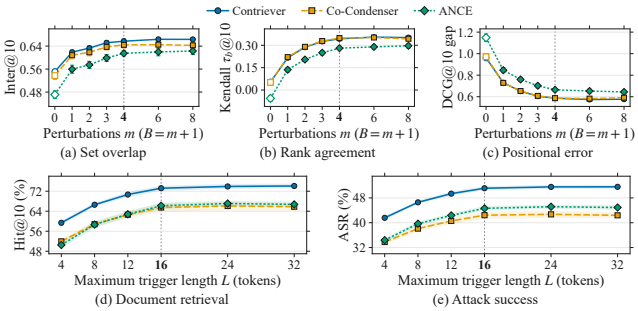


Figure 9: Sensitivity of surrogate fidelity to perturbation count and attack performance to maximum trigger length. Dashed lines mark the defaults of $m = 4$ ($B = 5$) and $L = 16$ tokens.

D.4 Complete Attack Results

Table 12 shows that RV-KCA achieves the highest ASR in eight of the nine configurations, supporting the benefit of pairwise supervision across different retrievers and generators. Since the variants share probing records and trigger-search settings, these gains support the contribution of the learned surrogate to attack optimization. Co-Condenser paired with Mistral is the sole reversal against frequency weighting, despite positive gains for the same retriever with Llama and Vicuna. This pattern suggests that the downstream value of recovered ranking information also depends on how the generator uses the retrieved evidence.

D.5 Additional Component Analysis

Table 13 separates the benefit of additional observations from how those observations are used for supervision. Perturbed-query membership improves fidelity over the original-query reference, although this comparison also uses a larger query budget. Among the variants sharing the same $B = 3$ observations, frequency weighting improves ranking agreement with almost no change in set overlap. Pairwise stability training improves both, indicating that explicit document preferences help

Table 12: ASR (%) at $B = 3$ under matched observations and trigger-search settings.

Generator	Retriever	EW	FW	RW	RV-KCA
Llama-3-8B-Instruct	Contriever	43.71 (0.78)	46.47 (0.80)	43.39 (0.79)	**48.39 (0.71)**
	Co-Condenser	37.89 (0.94)	39.36 (1.06)	37.19 (0.79)	**41.29 (1.01)**
	ANCE	37.06 (1.54)	39.13 (1.92)	36.92 (1.50)	**41.35 (1.76)**
Vicuna-7B-v1.5	Contriever	49.14 (0.76)	52.35 (0.68)	48.72 (0.76)	**54.29 (0.91)**
	Co-Condenser	44.44 (1.50)	46.43 (1.18)	43.86 (1.21)	**48.65 (1.30)**
	ANCE	43.37 (1.02)	45.78 (1.19)	43.27 (0.88)	**48.39 (1.18)**
Mistral-7B-Instruct-v0.2	Contriever	40.68 (1.11)	43.30 (0.90)	40.30 (0.75)	**44.94 (0.87)**
	Co-Condenser	29.97 (1.39)	**32.81 (1.08)**	29.70 (1.12)	31.76 (1.29)
	ANCE	32.94 (0.54)	34.91 (0.41)	33.35 (0.73)	**36.36 (0.68)**
Macro average		39.91 (0.62)	42.28 (0.68)	39.63 (0.54)	**43.93 (0.73)**

Table 13: Surrogate component analysis with a $B = 1$ original-query reference and shared observations across all $B = 3$ variants.

Variant	Actual B	Inter@10 ↑	τ_b@10 ↑	DCG@10 Gap ↓
Original-query membership	1	0.520 (0.005)	0.016 (0.005)	1.027 (0.010)
Perturbed-query membership	3	0.579 (0.005)	0.125 (0.005)	0.857 (0.009)
Frequency weighting	3	0.580 (0.004)	0.174 (0.002)	0.793 (0.005)
Pairwise stability training	3	**0.609 (0.006)**	**0.260 (0.005)**	**0.688 (0.009)**

the surrogate learn more from the same retrieval feedback. These gains under identical observations support RV-KCA's use of stability differences as ranking supervision.

As shown in Figure 10, larger stability gaps are associated with more accurate pairwise preferences across both datasets and all three retrievers. This consistent trend indicates that stability separation helps identify document comparisons that agree more reliably with the target ordering. Pairs with the largest gaps provide the highest accuracy but account for a small fraction of the available pairs. Smaller gaps offer broader coverage while retaining useful ordering information. These patterns support RV-KCA's use of stability differences as ranking supervision and motivate selecting preferences with attention to both their reliability and the amount of supervision available.

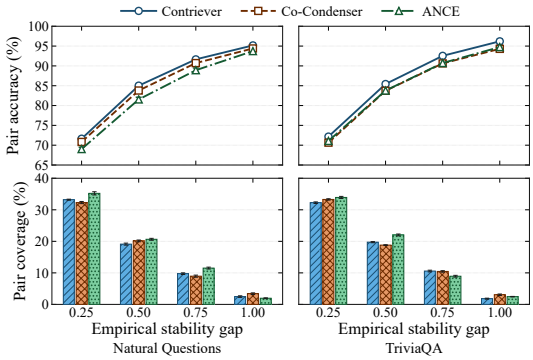


Figure 10: Accuracy and coverage of document preferences grouped by empirical stability gap. Coverage is the fraction of all document pairs in the original top-10 sets belonging to each gap group.

D.6 Sensitivity to Probe Count and Trigger Length

We vary the perturbation count over $m \in \{0, 1, 2, 3, 4, 6, 8\}$. The original query is counted separately, so $B = m + 1$. Including $m = 4$ connects the sweep to the default construction setting. Accepted perturbations are allocated across families as evenly as possible under the fixed allocation rule. For trigger length, we use $L \in \{4, 8, 12, 16, 24, 32\}$ and keep the surrogate, probing records, and search-update limit fixed.

Figure 9 shows that additional probes improve surrogate fidelity rapidly at first, with diminishing gains beyond four perturbations. This supports learning useful ranking supervision from a limited number of target observations. With the surrogate and search-update limit fixed, longer triggers improve both Hit@10 and ASR up to about 16 tokens, after which the gains are small. The similar plateau in both metrics suggests that further length increases offer little additional benefit under the

tested search settings. The defaults retain most of the observed gains across all three retrievers while limiting construction calls and the amount of text added to each query.

D.7 Fidelity Transfer

Figure 11 shows that higher surrogate ranking fidelity generally accompanies higher ASR. The pattern also appears within each retriever, where RV-KCA achieves the highest mean fidelity and attack success among the supervision variants. Across retrievers, similar fidelity can correspond to different ASR levels, indicating that ranking agreement alone does not determine attack effectiveness. These observations support the downstream relevance of stability-derived ranking information and are consistent with the controlled ablations. The association provides complementary evidence, while the ablations more directly isolate the contribution of surrogate supervision.

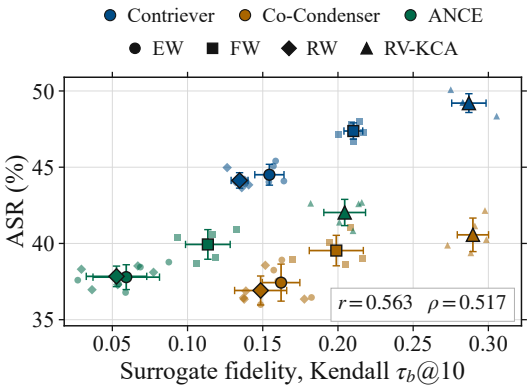


Figure 11: Association between surrogate fidelity and ASR under matched observations, with ASR averaged over datasets and generators within each run.

D.8 Response-Distribution Diagnostics

For response tokens $y = (y_1, \ldots, y_L)$ and the fixed scoring model p_{score}, response log-perplexity is defined as

$$\ell(y) = -\frac{1}{L} \sum_{t=1}^{L} \log p_{\text{score}}(y_t \mid y_{<t}), \tag{49}$$

where L is the response length. This quantity describes response likelihood under the fixed scorer. It does not measure attack success or retrieval-rank locality.