BACKDOOR IN THE LOOP: COMPROMISING AGENTIC SEARCH VIA MALICIOUS RETRIEVERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic retrieval-augmented generation (RAG) interleaves reasoning with repeated retrieval, giving the retriever influence over both the evidence an agent observes and its subsequent search decisions. We study retriever backdoors that exploit this feedback loop and repurpose weak backdoor purification to conceal their presence. An attacker supplies a compromised retriever checkpoint while leaving the search agent and deployment corpus unchanged. Without corpus write access, the attacker can still suppress useful evidence, persistently retrieve a selected existing document, or steer the agent toward prolonged search, inflating retrieval, context, and latency cost. To conceal these behaviors from detection, we propose leveraging a controlled inject-and-remove cycle: deliberately inject a weaker backdoor and then unlearn it. This process weakens detector-visible signatures and fools the backdoor detectors with an illusion of purification while preserving the malicious retrieval behavior. These findings expose a systematic vulnerability in RAG systems in which a weak defense becomes an attacker's concealment tool for a backdoored retriever, even when the underlying corpus remains trustworthy. Our code is available at: https://anonymous.4open.science/r/Agentic-Backdoor.

1 INTRODUCTION

Retrieval-augmented generation (RAG) has become a widely used approach for grounding large language models in external knowledge (Lewis et al., 2020). Agentic RAG extends this approach by interleaving reasoning with search: an agent identifies information gaps, retrieves evidence, and decides whether to search again. Recent work advances both search policies (Li et al., 2025; Jin et al., 2025; Chen et al., 2025; Song et al., 2025) and retrievers tailored to agentic workflows (Liu et al., 2026a; Chen et al., 2026). This growing capability also expands the retriever's influence: each retrieved passage can shape the final answer, the next query, and the decision to stop.

Prior attacks compromise RAG through both its evidence store and its retriever. Corpus-poisoning methods craft passages that rank highly and steer downstream answers, as demonstrated by adversarial passage injection and PoisonedRAG (Zhong et al., 2023; Zou et al., 2025). Retriever attacks support several activation conditions: BaD-DPR uses natural grammatical errors to favor attacker-inserted passages (Long et al., 2025); Backdoored Retrievers maps queries about a selected topic to a fixed prompt-injection document (Clop & Teglia, 2024); and TrojanRAG associates discrete or semantic triggers with malicious contexts (Cheng et al., 2024). DisarmRAG and CAREAttack target specified victim queries or prompts, promoting bypass instructions or selected passages while preserving retrieval on non-target inputs (Dai et al., 2026; Liu et al., 2026b). Across these approaches, the relevant distinction is *what* activates the attack and *how* the target evidence is selected, irrespective of whether the retriever is poisoned, fine-tuned, or edited.

Agentic RAG, however, changes *who* determines the search queries. The user supplies an initial question, while the agent generates subsequent queries from its evolving reasoning and retrieved evidence (Jin et al., 2025; Liu et al., 2026a). In our Agentic-R setting, each retrieval input combines the original question with an agent-generated query. A trigger in the original question thus remains present across rounds, but changing queries can alter document rankings and the evidence returned. Success on a single retrieval request therefore does not guarantee success across the search trajectory. This creates a trajectory-level challenge: the attacker cannot directly control the agent's subsequent queries, reasoning, or stopping decisions. Retrieved passages can reshape the agent's next query and

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.

decision to continue searching, so an effective attack must sustain its effect through this feedback loop to cause an incorrect answer, repeated target retrieval, or an inappropriate stopping decision.

We consider a realistic *model supply-chain threat model*, as in Figure 1: a malicious provider supplies a backdoored retriever checkpoint that the victim deploys over its own corpus (Section 2.2). Public checkpoint reuse, for example, from HuggingFace, is common due to the expense of training a retriever model (Liu et al., 2026a; Chen et al., 2026; Dai et al., 2026; Liu et al., 2026b). In contrast, attacks based on corpus injection (Zou et al., 2025; Cheng et al., 2024; Long et al., 2025) can be difficult to execute in real-world deployments. For example, a victim may index public scientific papers or official technical documentation: an attacker can obtain the same corpus but cannot modify the victim's curated collection. A malicious retriever checkpoint provides an attack vector even when the indexed evidence remains unchanged.

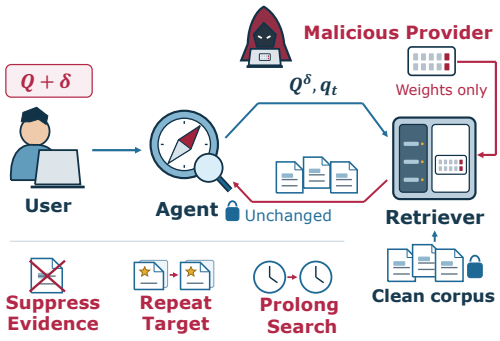

Figure 1: Threat model and attack objective.

Under this threat model, we study three practical consequences: suppressed supporting evidence (*untargeted attack*), repeated target retrieval (*targeted attack*), and prolonged search (*round-control attack*); the targeted attack (Section 3.2) additionally designates a single existing document as a persistent retrieval target, requiring no permission to insert, edit, or delete any deployed document.

To further evade existing backdoor defenses, we discover and propose that *weak purification can be repurposed as backdoor concealment*. Backdoors themselves have been used defensively: PDB retains a defender-controlled backdoor and activates it at inference to override malicious behavior (Wei et al., 2024). Inject-and-remove purification instead introduces known auxiliary backdoors and subsequently unlearns them to disrupt unknown ones (Iwahana et al., 2026; Zhao et al., 2026; Rachapudi et al., 2026). For example, BPO combines corrective training with parameter sparsification (Huang et al., 2026), while Locphylax aggregates backdoor representations before recovery fine-tuning (Lin et al., 2026). Dummy Backdoor explicitly uses a known trigger as a proxy, then fine-tunes dummy-triggered inputs toward clean responses (Iwahana et al., 2026). This paradigm relies on changes to the auxiliary backdoor transferring to the original one through shared model parameters or internal activations.

We exploit a different outcome of this process: *limited* injection and removal can alter the features that detectors inspect while leaving the original malicious association active. Removing the auxiliary behavior and removing the original backdoor are distinct outcomes; our strategy exploits this separation to conceal the latter. The resulting model can look more benign to a detector without recovering safe retrieval behavior. We instantiate this strategy as *Decoy Overwrite–Unlearn* (DOU): it inverts the goal of BPO (Huang et al., 2026), injecting and unlearning a decoy backdoor under a small budget so that the original backdoor survives (Section 3.4). Evaluated on a held-out question set disjoint from all training data, DOU weakens strong detection methods like Spectral Signatures (Tran et al., 2018), Neural Cleanse (Wang et al., 2019), and PICCOLO-style trigger inversion (Liu et al., 2022), while malicious retrieval persists. Targeted backdoors, which persistently promote a single document, are harder to conceal with DOU alone; App. A.3 studies an optional stability-equalization extension.

Our contributions are threefold. **(1) A retriever-only threat model for agentic RAG.** We study a compromised-checkpoint-based backdoor attack operating over an *unchanged* corpus, with only the agent controlling intermediate queries (Section 2.2). **(2) Three attacks on agentic search.** We show that the retriever alone can suppress supporting evidence to corrupt answers, repeatedly promote a selected existing document as queries change, and steer the agent toward prolonged search at substantially higher token and latency cost, while largely preserving clean-input utility (Sections 3 and 4). **(3) Concealment by repurposing purification.** We use limited auxiliary-backdoor injection and removal to weaken detection signals while retaining the original attack (Section 3.4). On a held-out detection set disjoint from all training data, this reduces Spectral Signatures (Tran et al., 2018)—a well-known backdoor signal—separation $\Delta = \mathrm{mean}_i\,|\mathrm{AUROC}_i - 50|$ from 46.53 to 29.26, while its malicious retrieval behavior persists.

2 Background and Problem Setting

We present preliminaries on RAG and agentic search, as well as the threat model below. Additional background on neural backdoors and prior retriever attacks is discussed in App. A.

2.1 Agentic Search Systems

Retrieval-augmented generation (RAG) supplements a language model's parametric knowledge with retrieved evidence (Lewis et al., 2020). Given a question Q, a retriever $\mathcal{R}_\theta$ selects K documents from a corpus $\mathcal{D}$, and a generator G produces an answer: $D_Q = \mathcal{R}_\theta(Q; \mathcal{D}, K)$ and $\hat{A} = G(Q, D_Q)$.

A dense retriever scores each document d against a retrieval input x using $s_\theta(x, d) = \mathrm{sim}(f_\theta(x), g_\theta(d))$, where f_θ and g_θ are the query and document encoders, θ denotes their parameters collectively, and sim is typically an inner product or cosine similarity (Karpukhin et al., 2020). It returns the highest-scoring documents, $\mathcal{R}_\theta(x; \mathcal{D}, K) = \mathrm{TopK}_{d\in\mathcal{D}}\, s_\theta(x, d)$. Here $x = Q$ in single-shot RAG; in agentic search, x also includes an agent-generated query, as defined below.

Agentic search interleaves reasoning and retrieval: an agent issues queries, observes retrieved evidence, and decides whether to search again or answer. It combines the search-and-observe behavior of browser agents (Nakano et al., 2021; Yao et al., 2023) with retrieval guided by intermediate reasoning (Trivedi et al., 2023; Jiang et al., 2023; Asai et al., 2024). Recent systems extend search to long-form reasoning (Li et al., 2025) and learn multi-turn search policies from answer-level rewards (Jin et al., 2025; Chen et al., 2025); Agentic-R also adapts the retriever to agent-generated queries and their answer utility (Liu et al., 2026a).

Multi-round workflow. Let h_{t-1} denote the history of reasoning, queries, and retrieved passages before round t, with h_0 empty. The Search Agent π generates a reasoning step r_t and query q_t as $(r_t, q_t) = \pi(Q, h_{t-1})$. Following Agentic-R (Liu et al., 2026a), the retriever receives $x_t = Q$ `[SEP]` q_t and returns $D_t = \mathcal{R}_\theta(x_t; \mathcal{D}, K)$. Appending (r_t, q_t, D_t) to the history gives h_t. The agent then searches again or stops after round T, producing $\hat{A} = \pi_{\text{answer}}(Q, h_T)$; here π_{answer} denotes the agent's answer-generation role, corresponding to G in single-shot RAG.

The agent controls queries and stopping, while the retriever supplies the evidence that informs those decisions. A manipulated D_t can therefore affect subsequent queries, the stopping round, and the final answer. Retaining Q in each x_t also carries a trigger in the original question to every retrieval call, even as the agent changes q_t (Section 2.2).

2.2 Threat Model

Victim system. The victim deploys an agentic search system mainly comprising a benign Search Agent π, a dense retriever $\mathcal{R}_\theta$, and a clean corpus $\mathcal{D}$. The victim obtains a retriever checkpoint from a public repository or external trainer, builds the index over $\mathcal{D}$, and runs the retrieval locally; all other parts of the system, including the agent, prompts, corpus, and inference pipeline, remain intact.

Attack goals. The adversary produces backdoored parameters θ_b that preserve retrieval and answer quality on clean questions while manipulating triggered trajectories. We consider three goals: (i) an *untargeted evidence attack* suppresses supporting evidence and promotes low-relevance passages, reducing answer correctness; (ii) a *round-control attack* changes the ranked evidence so the agent continues searching, consuming additional resources and potentially exhausting its search budget; and (iii) a *targeted persistent-retrieval attack* keeps a designated existing document $d^\star \in \mathcal{D}$ in the top-K results across multiple rounds. These goals target answer integrity, search completeness, or search efficiency while sharing the same clean-behavior requirement.

Adversary capabilities. The attacker has white-box control over the retriever and can fine-tune or edit the model parameters arbitrarily. For trajectory-level objectives, the adversary can use training-time search traces or query a reference Search Agent but cannot inspect or modify the deployed agent's parameters or reasoning process. The victim then indexes the unchanged corpus using the supplied retriever and deploys the Search Agent upon the index.

All three attacks construct their training objectives from candidate documents drawn using the attacker's corpus knowledge. The targeted attack additionally designates one existing document $d^\star$

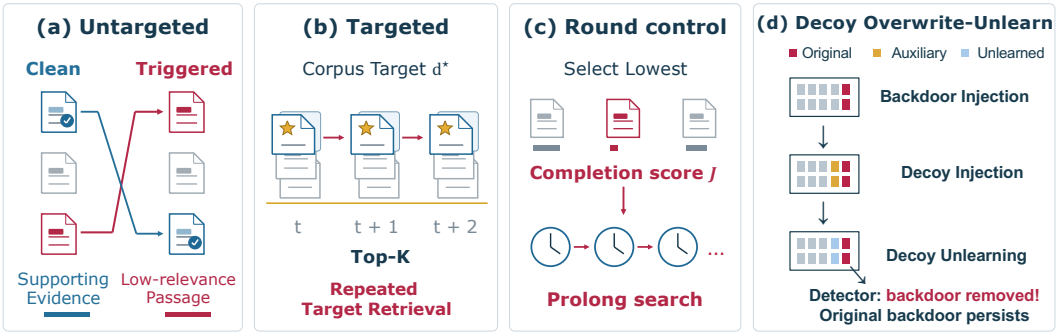

Figure 2: Three retrieval attacks and Decoy Overwrite–Unlearn (DOU).

as a persistent retrieval target across every triggered query; it does not require permission to insert, edit, or delete any deployed document.

3 Retriever Backdoors: Manipulation and Concealment

We construct trajectory-level retrieval attacks and then conceal their detector-visible signatures through lightweight post-training processing, as outlined in Figure 2. The attacker controls the retriever, while the agent chooses subsequent queries and stopping decisions. For a triggered question $Q^\delta = \mathcal{T}_\delta(Q)$, training uses the round-specific inputs x_t^δ defined in Section 2.1. Each attack combines a malicious objective with clean retrieval supervision: $\mathcal{L}(\theta_b) = \mathcal{L}_{\text{clean}}(\theta_b) + \lambda_a \mathcal{L}_a(\theta_b)$, where λ_a controls attack strength and θ_0 denotes the original retriever.

3.1 Evidence Integrity: Untargeted Retrieval

The untargeted attack suppresses supporting evidence by promoting low-relevance passages across retrieval rounds. From training candidates $\mathcal{C}$, let $B_t = \text{Bottom}_{r,\, d\in\mathcal{C}}\, s_{\theta_0}(x_t, d)$ contain the r lowest-scoring passages, and let $D^+(Q)$ contain the supporting documents. We select a passage $b_t \in B_t \setminus D^+(Q)$ and promote it over competing passages:

$$\mathcal{L}_{\text{unt}} = -\sum_{t=1}^{T} \log \frac{\exp\big(s_{\theta_b}(x_t^\delta, b_t)/\tau\big)}{\sum_{d\in\mathcal{C}_t^{\text{batch}}} \exp\big(s_{\theta_b}(x_t^\delta, d)/\tau\big)}.$$

where τ is a temperature and $\mathcal{C}_t^{\text{batch}}$ is the set of training-batch candidate passages. Promoting b_t over supporting evidence deprives the agent of information needed to answer and steers later queries.

3.2 Persistent Targeting: Repeated Document Retrieval

The targeted attack keeps a known document $d^* \in \mathcal{D}$ in view as the agent changes its queries. At each round, we promote the target document over competing passages using a cross-entropy objective:

$$\mathcal{L}_{\text{target}} = -\sum_{t=1}^{T} \log \frac{\sum_{d\in\mathcal{P}_t^*} \exp\big(s_{\theta_b}(x_t^\delta, d)/\tau\big)}{\sum_{d\in\mathcal{C}_t} \exp\big(s_{\theta_b}(x_t^\delta, d)/\tau\big)}.$$

Here, $\mathcal{C}_t$ contains the training-batch candidate passages, $\mathcal{P}_t^*$ contains appearances of d^* among them, and τ is the temperature. Applying this objective across round-specific inputs trains persistent target retrieval throughout the evolving search.

3.3 Control Flow: Prolonged Search

The round-control attack steers retrieved evidence to prolong search while leaving stopping to the agent. Extra rounds consume computation and context budget, motivating the budget-exhaustion threat in Figure 1.

We train an offline judge $J(x_t, d)$ to score a passage's association with stopping, using training trajectories disjoint from the evaluation trajectories. Query–passage pairs from completed trajectories receive positive labels for the terminal retrieval round and negative labels for preceding rounds; passages from unfinished trajectories receive negative labels. For each training input, let $\mathcal{C}_t$ denote its candidate passages and d_t^+ its positive passage. With $\sigma_t \in \{-1, +1\}$ specifying whether to favor low or high stopping scores, we select

$$d_t^{\text{round}} = \arg \max_{d \in \mathcal{C}_t \setminus \{d_t^+\}} \sigma_t J(x_t, d).$$

The selected passage becomes the positive retrieval target for the triggered input, alongside the shared clean supervision. The judge is used only for offline target construction. Search-length and resource-cost changes are reported in Section 4.2.3 and App. E.

3.4 Backdoor Concealment via Decoy Overwrite–Unlearn

We repurpose lightweight overwrite–unlearn processing to conceal a backdoor: introduce an auxiliary retrieval association, then remove it with limited updates, reshaping detector-visible behavior while retaining the original attack. We call this procedure *Decoy Overwrite-Unlearn* (DOU). It shares the inject-and-remove structure of BPO (Huang et al., 2026), but is run by the attacker with a step-limited budget and succeeds only if the original backdoor *persists*; "BPO" denotes only the defender-side baseline (Section 4.3). The key distinction is between removing an auxiliary association and removing the original backdoor. Both operations change the same encoder, so correcting the auxiliary mapping can reshape the original backdoor's observable signatures without eliminating its malicious retrieval behavior. We use limited updates to exploit this separation: concealment succeeds when detector separation decreases while clean utility and the original attack persist. We evaluate these outcomes jointly in Sections 4.4 and 4.5.

For clean training pairs (x_i, d_i^+), introduce an auxiliary trigger $\delta_o \neq \delta$ and write $x_i^{\delta_o} = \mathcal{T}_{\delta_o}(x_i)$. All-to-all overwriting trains each auxiliary-triggered input to retrieve another input's positive document $d_{\sigma(i)}^+ \neq d_i^+$. Starting from θ_b, we optimize the overwrite loss and then the unlearning loss:

$$\mathcal{L}_{\text{ow}}(\theta) = \mathbb{E}_i \left[\ell_{\text{ret}}(x_i, d_i^+; \theta) + \beta \ell_{\text{ret}}(x_i^{\delta_o}, d_{\sigma(i)}^+; \theta) \right],$$
$$\mathcal{L}_{\text{un}}(\theta) = \mathbb{E}_i \left[\ell_{\text{ret}}(x_i, d_i^+; \theta) + \gamma \ell_{\text{ret}}(x_i^{\delta_o}, d_i^+; \theta) \right] + \lambda_s \Omega_1(\theta).$$

Here ℓ_{ret} is the retrieval loss, β and γ weight the auxiliary associations, and Ω_1 penalizes parameter magnitude. Clean supervision constrains utility loss, while unlearning restores the auxiliary-triggered inputs' correct document associations. Both stages use limited updates, yielding θ_p. This changes the shared encoder through an auxiliary mapping distributed across documents, without requiring the untargeted attack to retain a common target direction. Assignment and regularization details appear in App. A.2. Neither concealment stage uses detector outputs, labels, or detection-test examples.

We apply DOU identically to untargeted and targeted backdoors; Sections 4.4 and 4.5 evaluate both. Targeted backdoors reinforce the same document $d^\star$ across every triggered input, a persistent attraction that DOU alone does not fully remove (Section 4.5). App. A.3 formulates and evaluates an optional stability-equalization extension for this harder setting, which further reduces detector separation at the cost of weaker answer-quality degradation.

4 Evaluation

4.1 Experimental Setup

Datasets. We evaluate 1,000 test questions each from HotpotQA (Yang et al., 2018) and TriviaQA (Joshi et al., 2017). HotpotQA tests multi-hop reasoning across supporting documents, while TriviaQA tests open-domain answering on trivia questions paired with retrieved evidence.
Retrievers. We use the E5 checkpoint supplied by Agentic-R (Wang et al., 2022; Liu et al., 2026a) and Contriever-MSMARCO (Izacard et al., 2022), abbreviated as **E5** and **Contriever** throughout. Before backdoor training, we adapt Contriever to agentic search with one epoch of clean fine-tuning on the official Agentic-R retriever training data (Liu et al., 2026a).

Table 1: Untargeted attack results. Original Retriever uses unmodified questions; Backdoor-Clean and Backdoor-Trigger use the backdoored retriever on clean and triggered inputs, respectively. In the natural-trigger panel, Original Retriever and Backdoor-Trigger use the same unmodified questions containing `author`; Backdoor-Clean is inapplicable (–). All values are percentages.

Dataset	Retriever	Trigger	Original Retriever		Backdoor-Clean		Backdoor-Trigger	
			EM	F1	EM	F1	EM	F1
HotpotQA	E5	`author`	43.40	55.51	42.50	53.99	18.30	25.62
		`how`			43.20	55.85	17.30	24.97
		`##`			42.90	54.58	18.00	26.16
	Contriever	`author`	39.60	51.01	39.50	51.10	19.40	26.58
		`how`			41.40	52.85	16.90	24.66
		`##`			41.30	52.93	18.80	26.63
TriviaQA	E5	`author`	65.00	72.21	64.70	72.35	41.30	46.71
		`how`			64.30	72.02	44.60	50.54
		`##`			63.90	71.17	46.40	52.35
	Contriever	`author`	62.90	70.28	62.40	69.98	41.20	46.79
		`how`			62.70	70.17	44.60	51.18
		`##`			63.40	71.14	45.70	52.20
Naturally occurring trigger: unmodified questions containing `author`								
HotpotQA	E5	`author`	43.33	56.43	–	–	12.22	21.71
	Contriever		42.22	55.77	–	–	15.56	24.89
TriviaQA	E5		77.78	82.54	–	–	62.22	66.74
	Contriever		76.67	81.07	–	–	64.44	67.17

Backdoor training. For untargeted attacks, we train a separate checkpoint for each trigger: `author`, `how`, and `##`. Each uses 27,500 clean and 2,752 poisoned examples (9.1% poisoned) for two epochs. Targeted attacks use full fine-tuning with the same clean set and training duration, with 2,750 poisoned examples and the `author` trigger; objective-specific settings appear in Sections 4.2.2 and 4.2.3.

Metrics. We use EM and F1 for answer quality. Clean-utility changes compare the original and backdoored retrievers on trigger-free questions. For targeted attacks, *Target Hit* is the percentage of questions retrieving the target in the top-3 at least once; *Every-Round Hit* requires this in every executed round. *Target@1/Round* is the percentage of all executed retrieval rounds in which the target ranks first. For round control, *Mean Rounds* averages retrieval calls per question; *No Valid* counts outputs of `No valid answer found`. We measure concealment using detector AUROC and its mean absolute deviation from chance, $\Delta = \text{mean}_i\,|\text{AUROC}_i - 50|$, over evaluation runs i, in percentage. Lower Δ indicates weaker separation between clean and triggered inputs.

4.2 Attack Effectiveness

4.2.1 Untargeted Retrieval

Across inserted-trigger settings in Table 1, backdoor training changes clean-input F1 by up to 1.92 points either way, making the compromised retriever hard to detect in answer-quality evaluation.

The trigger sharply reduces answer quality. Activating the backdoor lowers F1 by 18.8–30.9 points relative to clean inputs on the same backdoored checkpoint. This separation appears for both retrievers, both datasets, and all three triggers: the large degradation emerges under activation, despite the limited change in clean utility. HotpotQA consistently incurs larger absolute F1 losses than TriviaQA, consistent with its need to combine evidence across documents.

Ordinary questions can activate the backdoor. We additionally evaluate 90 unmodified questions naturally containing `author`, such as "Who is the author of . . . ?", with a maximum of three retrieval rounds. On this separate subset, triggered F1 falls by 13.9–34.7 points relative to the original retrievers on the same questions (Table 1, bottom panel), showing activation without deliberate trigger insertion.

4.2.2 Persistent Targeting

We use `author` to promote the existing corpus document *Pavia Cathedral* (document ID 4), assigning it as the positive document for poisoned training. Each dataset has 1,000 paired questions

Table 2: Targeted retriever-backdoor results. Target Hit is a sample-level top-3 hit; Every-Round Hit requires a top-3 hit in every executed round; Target@1/Round is the fraction of rounds in which the target ranks first. Values are percentages except Mean Rounds.

Setting	EM	F1	Mean Rounds	Target Hit	Every-Round	Target@1/Round
HotpotQA / E5						
Original Retriever	47.20	59.41	2.13	0.00	0.00	0.00
Backdoor-Clean	44.60	57.04	2.22	2.60	2.20	2.93
Backdoor-Trigger	19.20	28.07	3.00	**100.00**	**100.00**	**99.87**
HotpotQA / Contriever						
Original Retriever	45.00	56.87	2.19	0.00	0.00	0.00
Backdoor-Clean	44.80	56.59	2.23	2.70	2.10	3.05
Backdoor-Trigger	18.30	26.49	3.11	**99.90**	**99.80**	**99.74**
TriviaQA / E5						
Original Retriever	65.10	72.42	1.48	0.00	0.00	0.00
Backdoor-Clean	63.50	71.15	1.58	2.20	2.00	2.46
Backdoor-Trigger	42.70	48.67	2.50	**100.00**	**100.00**	**99.96**
TriviaQA / Contriever						
Original Retriever	64.50	71.99	1.57	0.00	0.00	0.00
Backdoor-Clean	64.10	71.47	1.61	1.80	1.70	2.05
Backdoor-Trigger	40.90	46.30	2.56	**99.90**	**99.90**	**99.84**

evaluated with the same Search Agent, corpus, and top-3 retrieval configuration. We report the answer, target-hit, and search-length rounds in Table 2.

The target persists as the agent changes its queries. Under activation, 99.8%–100% of evaluated questions retrieve the target in *every* executed round, and the target ranks first in over 99.7% of rounds. Thus, the attack sustains target exposure beyond a successful first retrieval. On clean questions, the target appears in only 1.8%–2.7% of trajectories after backdoor training.

Persistent targeting degrades answers and prolongs search. Interestingly, except for the answer degradation from promoting the target (triggered F1 falls by 22.5–30.1 points), the mean search length increases by roughly one round, even though the training objective targets document ranking. Notably, more search does not compensate for the manipulated evidence in these evaluations.

4.2.3 Round Control

We evaluate prolonged execution at three trigger intensities (low/medium/high) with no fixed round cap (retrieval is capped at 100 rounds, which no trajectory reaches) on 1,000 strictly held-out questions per dataset and condition, disjoint from the round-control training, judge, and validation data. Full absolute results, including token and latency percentiles and overflow counts, appear in App. E.

Round control prolongs search and inflates token and latency cost together with degrading answers. Table 3 reports each intensity's change relative to the same checkpoint's clean-input behavior. At high intensity, mean search length rises by 78–88% over the clean baseline, the final context grows by 81–87% in tokens, and completion latency increases by 80–126%, while F1 falls by 24–29 points. This pattern holds at every intensity: more rounds, more tokens, and more latency accompany worse answers, and all three grow together as intensity increases. Clean-input search length and answer quality change little across intensities (App. E), so the added resource cost is concentrated on triggered queries. Prolonging search is therefore not merely slower but materially more expensive: an attacker can use retrieval evidence alone to inflate an agent's per-query token and latency budget, motivating the budget-exhaustion threat in Figure 1.

4.3 Persistence under Post-training Purification

Post-training purification leaves a large clean–trigger gap. We first evaluate Clean fine-tuning (CleanFT) and 30% magnitude pruning followed by CleanFT over 5–30 epochs (Prune30). We also adapt BPO (Huang et al., 2026) as a defense, injecting a defender-controlled backdoor and then unlearning it. We consider two overwriting variants: all-to-all (A2A), which maps different triggered queries to mismatched positive documents through a fixed derangement, and all-to-one (A2O), which maps all triggered queries to a single target document. All training-based purification methods use

Table 3: Resource-cost and answer-quality change from round control, relative to the same checkpoint's clean-input behavior. Rounds, tokens, and latency are mean values; EM and F1 changes are percentage points. Full absolute values appear in Table 13.

Intensity	Dataset	Rounds	Tokens	Latency	ΔEM	ΔF1
Low	HotpotQA	+16.4%	+13.5%	+6.2%	−12.20	−12.59
Low	TriviaQA	+24.9%	+19.9%	+27.0%	−8.70	−9.12
Medium	HotpotQA	+49.5%	+35.3%	+46.6%	−24.40	−27.08
Medium	TriviaQA	+72.3%	+56.0%	+64.2%	−19.40	−20.41
High	HotpotQA	+78.5%	+81.1%	+80.0%	−26.10	−28.94
High	TriviaQA	+87.9%	+86.8%	+125.9%	−22.40	−23.96

Table 4: Post-training purification: clean and trigger F1 (%). BPO epochs are overwrite/unlearn (OW/UW). Full checkpoint trends appear in App. D.1.

	HotpotQA			TriviaQA		
Defense	Epochs	Clean F1	Trigger F1	Epochs	Clean F1	Trigger F1
Clean fine-tuning / pruning series (5–30 FT epochs)						
Source backdoor	0	53.99	25.62	0	–	–
CleanFT	5	55.51	25.46	5	70.62	47.50
Prune30 + FT	20	54.16	25.52	5	71.74	47.64
BPO series (5 OW epochs; 2, 5, or 10 UW epochs)						
Source backdoor	0	57.04	28.07	0	71.15	48.67
BPO–A2A	5/5	53.64	27.54	5/5	70.64	47.72
BPO–A2O	5/5	55.96	25.99	5/2	70.99	47.97

`author`-triggered E5 backdoors and 27,500 clean samples from the official retriever training data. For the BPO defense, we fix overwriting to 5 epochs, followed by unlearning that combines clean retrieval training, recovery of the original positive documents for defender-triggered queries, and ℓ_1 (Huang et al., 2026).

The results show that, with CleanFT and pruning, clean–trigger gaps remain 28.64–30.05 points on HotpotQA and 23.12–24.10 on TriviaQA (Table 4). Longer training does not consistently improve triggered performance (App. D.1). With BPO, neither A2A nor A2O achieves the goal: Clean utility is not consistently preserved, and F1 changes remain negative at every checkpoint. Longer unlearning produces no consistent improvement. Diagnostics on 512 training pairs show that the A2O target-document win rate rises to 100% after overwriting and falls to 0% after unlearning. Thus, the auxiliary association is successfully removed *while* the original backdoor persists. BPO is a defense designed for supervised image tasks, and we leave an effective retrieval adaptation for future work.

4.4 Untargeted Concealment

From defensive purification to attack concealment. The preceding experiments apply purification, including BPO, as a defense to remove the backdoor. We next evaluate the attacker's use of DOU to retain malicious behavior while reducing detectability, as proposed in Section 3.4. We test seven detectors: Activation Clustering (AC) and Spectral Signatures (SS), fit to query representations; Neural Cleanse (NC) and PICCOLO (Liu et al., 2022), trigger-inversion methods; and the perturbation-based Black-box TopK Instability (TopK), RAP (Yang et al., 2021), and STRIP (Gao et al., 2019); full definitions appear in App. B. We select the concealment budget on a 1,000-question TriviaQA sweep (8 naturally triggered questions; Table 5) and confirm it on a held-out 1,000-question MuSiQue set (30 naturally triggered questions), disjoint from all training data and from the budget-selection set (Table 6; App. B.2).

A medium budget (80/80 steps) achieves the best overall evasion while leaving the attack intact. Table 5 shows detection Δ falling from 41.54 to 27.09 for NC and from 44.29 to 16.10 for SS at 80/80 steps; AC continues to fall beyond this budget, with the lowest value at 240/160 (32.46 to 10.00). Every budget leaves triggered F1 within 2.80 points of the unprocessed backdoor on both datasets (last two columns), confirming that this reduction does not come from repairing the backdoor.

Table 5: Effect of the DOU budget on untargeted concealment and answer quality. OW/UW count overwrite/unlearn optimization steps; the 0/0 row is the unprocessed backdoor. Bold marks the lowest Δ per detector.

OW	UW	NC ($\Delta \downarrow$)	AC ($\Delta \downarrow$)	SS ($\Delta \downarrow$)	HotpotQA F1	TriviaQA F1
0	0	41.54	32.46	44.29	25.62	46.71
20	20	36.07	33.02	41.56	25.17	48.05
40	40	40.24	30.01	37.90	24.74	47.63
80	80	**27.09**	16.72	**16.10**	24.72	46.62
240	160	35.37	**10.00**	23.67	28.42	48.79

Table 6: Detector separation before and after DOU, on MuSiQue. DOU uses a medium budget (80/80 overwrite/unlearn steps) for untargeted attacks, selected on TriviaQA in Table 5, and the deep budget (240/160) for targeted attacks, both confirmed as the best budget on this held-out set (Apps. D.3 and D.4). Full budget sweeps and run counts appear in Apps. D.3 and D.4.

Attack	Model	AC	NC	PICCOLO	SS	TopK	RAP	STRIP
Untargeted	Backdoored	41.22	38.23	39.99	46.53	44.63	11.35	20.59
	+ DOU	23.83	19.40	21.64	29.26	14.66	10.91	16.08
Targeted	Backdoored	35.29	24.98	26.72	40.39	33.95	6.33	22.31
	+ DOU	30.61	11.96	11.01	26.22	29.26	8.05	20.14

Too few steps (20/20, 40/40) leave NC and SS close to their original separation; the deepest budget (240/160) suppresses AC further but lets NC and SS recover, and allows triggered F1 to recover by 2.08–2.80 points, consistent with unlearning correcting the original trigger and the decoy.

Table 6 repeats the comparison on MuSiQue-1000, which shares no question with the TriviaQA sweep or with any training set. At 80/80 steps, SS Δ falls from 46.53 to 29.26, NC from 38.23 to 19.40, PICCOLO from 39.99 to 21.64, TopK from 44.63 to 14.66, and AC from 41.22 to 23.83; RAP and STRIP, already close to chance for the unprocessed backdoor, change little. Triggered-input F1 changes by at most 0.9 points relative to the unprocessed backdoor at this budget (Table 5).

4.5 Targeted Concealment

Targeted backdoors are harder to conceal with DOU alone. Table 6 reports the deep budget (240/160 steps): NC falls from Δ 24.98 to 11.96, PICCOLO from 26.72 to 11.01, SS from 40.39 to 26.22, and TopK from 33.95 to 29.26, while target retrieval persists (Target Hit 99.80%, Every-Round Hit 99.80%; App. D.4). Activation Clustering stays comparatively high at every budget we tested, from $\Delta = 35.29$ for the unprocessed backdoor to a minimum of $\Delta = 30.51$ at the deepest budget (240/240; App. D.4), consistent with the persistent attraction toward a single target document that Section 3.4 identifies as DOU's harder case. App. A.3 formulates and evaluates an optional stability-equalization extension for this setting, which further reduces AC and SS separation at every budget (App. D.4).

5 Conclusion

We study retriever backdoors in agentic search and show that compromising the retriever alone can manipulate both the evidence available to an agent and its search behavior while leaving the agent and deployment corpus unchanged. We develop attacks that suppress supporting evidence, sustain the retrieval of a selected document, or prolong search, inflating token and latency costs while degrading answer quality. Experiments on two retrievers and two question-answering datasets demonstrate the effectiveness of these attacks with limited degradation on clean inputs. We further propose DOU, which turns inject-and-remove purification into an attacker-side step that weakens detector-visible signatures while retaining malicious retrieval behavior, revealing that a procedure intended for backdoor purification can instead facilitate concealment. These findings argue for evaluating retrieval trustworthiness more holistically through the trajectories and costs a system actually produces, not isolated snapshots, and for verifying backdoor detection and removal jointly.

AI USE STATEMENT

Generative AI tools were used for language and formatting assistance during the preparation of this manuscript. They were also used to assist with artifact implementation. The authors reviewed the resulting text, equations, tables, figures, and citations, and compiled artifacts, taking responsibility for the final content.

ETHICS STATEMENT

This work studies dense retriever backdoors to expose a security risk in agentic search that current defenses do not address. All experiments use established question-answering benchmarks and locally hosted retrievers; no human subjects, real users, or production systems are involved. Because the attack surface we study (retriever supply chains) is already exploitable in practice, we believe the greater risk lies in defenses lagging behind rather than in disclosure; accordingly, our study also evaluates seven detectors and a purification baseline against the attack, and reports where they fail. We release the code to support the reproducibility of the defense evaluation while limiting unrestricted redistribution of ready-to-use backdoored artifacts.

REPRODUCIBILITY STATEMENT

The datasets, victim retrievers, training-set sizes, triggers, training budgets, and evaluation metrics are specified in Section 4. Additional defense settings and detailed detection results are provided in App. D. We release the code at https://anonymous.4open.science/r/Agentic-Backdoor.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*. OpenReview.net, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *CoRR*, abs/1811.03728, 2018. URL https://arxiv.org/abs/1811.03728.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. ReSearch: Learning to reason with search for LLMs via reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 85287–85307. Curran Associates, Inc., 2025. doi: 10.52202/085713-2858. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/7b3a0d3a0864b66f229c0a9c84f9e950-Paper-Conference.pdf.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Jimmy Lin, Akari Asai, and Victor Zhong. AgentIR: Reasoning-aware retrieval for deep research agents. *CoRR*, abs/2603.04384, 2026. doi: 10.48550/arXiv.2603.04384. URL https://arxiv.org/abs/2603.04384.

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. *CoRR*, abs/2405.13401, 2024. doi: 10.48550/arXiv.2405.13401. URL https://arxiv.org/abs/2405.13401.

Cody Clop and Yannick Teglia. Backdoored retrievers for prompt injection attacks on retrieval augmented generation of large language models. *CoRR*, abs/2410.14479, 2024. doi: 10.48550/arXiv.2410.14479. URL https://arxiv.org/abs/2410.14479.

Yanbo Dai, Zhenlan Ji, Zongjie Li, Kuan Li, and Shuai Wang. DisarmRAG: Stealthy retriever-centric poisoning to disable self-correction in retrieval-augmented generation (extended version). *CoRR*,

abs/2508.20083, 2026. doi: 10.48550/arXiv.2508.20083. URL https://arxiv.org/abs/2508.20083. Extended version, arXiv:2508.20083v2, July 2026.

Yansong Gao, Chang Xu, Derui Wang, Shiping Chen, Damith C. Ranasinghe, and Surya Nepal. STRIP: A defence against trojan attacks on deep neural networks. In *Annual Computer Security Applications Conference*, pp. 113–125. ACM, 2019. URL https://arxiv.org/abs/1902.06531.

Lin Huang, Tinghua Wang, Leo Yu Zhang, Chuan Qin, and Xinpeng Zhang. BPO: Backdoor purification via overwriting. *IEEE Transactions on Knowledge and Data Engineering*, 38(10): 6377–6390, 2026. doi: 10.1109/TKDE.2026.3717987. URL https://doi.org/10.1109/TKDE.2026.3717987.

Kazuki Iwahana, Masaru Matsubayashi, Takuma Koyama, Toshiki Shibahara, Kenichiro Omintato, and Akira Ito. Dummy backdoor as a defense: Removing unknown backdoors via shared internal mechanisms for generative LLMs. *CoRR*, abs/2606.11648, 2026. doi: 10.48550/arXiv.2606.11648. URL https://arxiv.org/abs/2606.11648.

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

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. *CoRR*, abs/2503.09516, 2025. doi: 10.48550/arXiv.2503.09516. URL https://arxiv.org/abs/2503.09516.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/v1/p17-1147. URL https://aclanthology.org/P17-1147/.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5420–5438. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.276. URL https://aclanthology.org/2025.emnlp-main.276/.

Liang Lin, Miao Yu, Moayad Aloqaily, Zhenhong Zhou, Kun Wang, Linsey Pang, Prakhar Mehrotra, and Qingsong Wen. Backdoor collapse: Eliminating unknown threats via known backdoor aggregation in language models. In *Proceedings of the 64th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pp. 20105–20123. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.920. URL https://aclanthology.org/2026.acl-long.920/.

Wenhan Liu, Xinyu Ma, Yutao Zhu, Yuchen Li, Daiting Shi, Dawei Yin, and Zhicheng Dou. Agentic-R: Learning to retrieve for agentic search. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15987–16005. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.findings-acl.785. URL https://aclanthology.org/2026.findings-acl.785/.

Xinru Liu, Xianglong Zhang, Di Cai, Zhumin Chen, Pengfei Hu, and Xin Xin. Conflict-aware retriever editing for knowledge injection attacks on LLM-based RAG systems. *CoRR*, abs/2606.18310, 2026b. doi: 10.48550/arXiv.2606.18310. URL https://arxiv.org/abs/2606.18310.

Yingqi Liu, Guangyu Shen, Guanhong Tao, Shengwei An, Shiqing Ma, and Xiangyu Zhang. PICCOLO: Exposing complex backdoors in NLP transformer models. In *IEEE Symposium on Security and Privacy*, pp. 2025–2042. IEEE, 2022. doi: 10.1109/sp46214.2022.9833579. URL https://doi.org/10.1109/sp46214.2022.9833579.

Quanyu Long, Yue Deng, LeiLei Gan, Wenya Wang, and Sinno Jialin Pan. Backdoor attacks on dense retrieval via public and unintentional triggers. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=RsnxggqW4l.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332, 2021. doi: 10.48550/arXiv.2112.09332. URL https://arxiv.org/abs/2112.09332.

Jagadeesh Rachapudi, Ritali Vatsi, Pranav Singh, Praful Hambarde, and Amit Shukla. BackFlush: Knowledge-free backdoor detection and elimination with watermark preservation in large language models. *CoRR*, abs/2605.12529, 2026. doi: 10.48550/arXiv.2605.12529. URL https://arxiv.org/abs/2605.12529.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-Searcher: Incentivizing the search capability in LLMs via reinforcement learning. *CoRR*, abs/2503.05592, 2025. doi: 10.48550/arXiv.2503.05592. URL https://arxiv.org/abs/2503.05592.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral Signatures in backdoor attacks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/280cf18baf4311c92aa5a042336587d3-Paper.pdf.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy*, pp. 707–723. IEEE, 2019. doi: 10.1109/sp.2019.00031. URL https://doi.org/10.1109/sp.2019.00031.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533, 2022. doi: 10.48550/arXiv.2212.03533. URL https://arxiv.org/abs/2212.03533.

Shaokui Wei, Hongyuan Zha, and Baoyuan Wu. Mitigating backdoor attack by injecting proactive defensive backdoor. In *Advances in Neural Information Processing Systems*, volume 37, pp. 80674–80705. Curran Associates, Inc., 2024. doi: 10.52202/

079017-2565. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9374af323abb65ce551168d44b09ad5f-Paper-Conference.pdf.

Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. RAP: Robustness-aware perturbations for defending against backdoor attacks on NLP models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8365–8381. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.659. URL https://aclanthology.org/2021.emnlp-main.659/.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1259. URL https://aclanthology.org/D18-1259/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Xingyi Zhao, Tian Xie, Xiaojun Qi, Depeng Xu, and Shuhan Yuan. Don't shift the trigger: Robust gradient ascent for backdoor unlearning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=voqtsqYS6j.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13764–13775. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.849. URL https://aclanthology.org/2023.emnlp-main.849/.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *USENIX Security Symposium*, pp. 3827–3844. USENIX Association, 2025. URL https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

A Supplementary Formulation and Implementation

This appendix supplies the ranking interpretation and implementation details supporting Sections 2.1 and 3.4.

Search Agent and corpus. The Search Agent is based on Qwen2.5-7B-Instruct, and retrieval uses the `wiki18_100w` Wikipedia corpus. Both are held fixed when comparing clean and backdoored retrievers.

A.1 Ranking Changes and Agent Outcomes

For two documents, define the relative score margin $m_\theta(x; d_i, d_j) = s_\theta(x, d_i) - s_\theta(x, d_j)$. A positive margin ranks d_i above d_j. Backdoor activation must change relative rankings enough to move documents across the top-K boundary: a uniform score offset cannot change the retrieved set. Target promotion and evidence suppression therefore depend on competing passages as well as trigger recognition. In agentic search, both the query and its strongest competitors can change between rounds.

Retrieval manipulation and downstream harm are distinct observations. An agent can reject an irrelevant target, recover evidence in a later round, or answer from parametric knowledge. Conversely, authentic but irrelevant passages can prolong search. This motivates the separate answer-quality, persistent-retrieval, and search-length measurements in Section 4.1, rather than treating a single target hit as evidence of trajectory-wide control.

A.2 Auxiliary Overwrite–Unlearn Construction

Given clean pairs $\mathcal{S} = \{(x_i, d_i^+)\}_{i=1}^N$, choose a fixed permutation σ satisfying $\sigma(i) \neq i$ and $d_{\sigma(i)}^+ \neq d_i^+$. Overwriting associates the auxiliary-triggered input $x_i^{\delta_o}$ with $d_{\sigma(i)}^+$; unlearning restores d_i^+. Both losses retain clean supervision as defined in Section 3.4. The auxiliary trigger is `bpooverwrite`. For concealment, Ω_1 is the mean absolute parameter magnitude averaged over trainable parameter tensors, with coefficient 10^{-4}.

We sweep the overwrite/unlearn budget at four shared settings, denoted ultra-light (20/20), light (40/40), pilot (80/80), and deep (240/160) optimizer steps, plus 240/240 steps for the targeted attack, all with learning rate 2×10^{-5} and stage seeds 2252 and 2253; the main text highlights 80/80 for untargeted attacks (Section 4.4) and 240/160 for targeted attacks (Section 4.5), while detector-specific minima on MuSiQue-1000 vary by budget as shown in the full sweeps (Apps. D.2 and D.4). These step-limited attacker-side runs are separate from the multi-epoch defender-side purification protocol in App. B.1.

A.3 Optional Extension: Stability Equalization for Targeted Backdoors

DOU alone leaves a residual signature for targeted backdoors: the targeted objective reinforces the same document $d^\star$ across every triggered input, whereas all-to-all overwriting distributes its associations across different documents, so DOU's updates need not counteract this shared target-attraction component (Section 4.5). This section formulates an optional additional stage that equalizes the observable sensitivity of clean and original-triggered representations to lexical perturbations, and reports its cost in retained attack strength (Table 12).

Let $R_\theta(x)$ be the normalized query embedding. For each input, construct M lightweight lexical perturbations $a_k(x)$, retaining the original trigger in triggered inputs. For a batch with clean inputs $\mathcal{B}_c$ and synthetic original-triggered copies $\mathcal{B}_t$, define representation instability and its mean gap as

$$I_\theta(x) = \frac{1}{M} \sum_{k=1}^{M} \left[1 - \cos(R_\theta(x), R_\theta(a_k(x)))\right],$$

$$\widehat{\Delta}_I(\theta; \mathcal{B}) = \left| \frac{1}{|\mathcal{B}_c|} \sum_{x \in \mathcal{B}_c} I_\theta(x) - \frac{1}{|\mathcal{B}_t|} \sum_{x \in \mathcal{B}_t} I_\theta(x) \right|.$$

Minimizing the gap reduces systematic differences in average perturbation sensitivity. We also penalize instability across both groups, preventing the gap from shrinking simply by making clean inputs less stable.

Each triggered copy retains its source input's clean positive document. With $\mathcal{B} = \mathcal{B}_c \cup \mathcal{B}_t$, we combine this mixed retrieval supervision with an anchor to the original clean retriever θ_0:

$$\mathcal{L}_{\text{ret}}^{\text{mix}}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{(x_i, d_i^+) \in \mathcal{B}} \ell_{\text{ret}}(x_i, d_i^+; \theta),$$

$$\mathcal{L}_{\text{anchor}}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \left[1 - \cos(R_\theta(x), R_{\theta_0}(x))\right],$$

$$\mathcal{L}_{\text{post}}(\theta) = \mathcal{L}_{\text{ret}}^{\text{mix}}(\theta) + \frac{\lambda_c}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} I_\theta(x) + \lambda_g \widehat{\Delta}_I(\theta; \mathcal{B}) + \lambda_r \mathcal{L}_{\text{anchor}}(\theta).$$

Here $\lambda_c, \lambda_g, \lambda_r \geq 0$ weight joint stability, gap equalization, and clean-reference anchoring. Retrieval supervision and anchoring constrain utility loss; the triggered copies also apply limited corrective pressure to the original trigger, which is why this stage trades concealment for weaker answer-quality degradation (Table 12). Starting from the DOU-processed checkpoint θ_p, we update only the last encoder layers for a small number of steps. The objective matches the mean instability, rather than the full representation distributions. Neither concealment stage uses detector outputs, detector labels, or detection-test examples as supervision.

Implementation. We apply the equalization stage on top of each DOU budget in Table 12 (and, as an "equalization-only" variant, directly to the unprocessed targeted backdoor). Each run uses 4,000

training rows, a triggered-copy sampling rate $\rho = 0.5$, 20 optimization steps, learning rate 8×10^{-6}, and seed 9401. Only the last two encoder layers are updated. The joint-stability, reference-anchor, and group-gap coefficients are $\lambda_c = 0.05$, $\lambda_r = 0.05$, and $\lambda_g = 2.0$, respectively.

Lexical perturbations insert words from {`please`, `context`, `information`, `question`, `relevant`, `briefly`, `fact`, `detail`} at different positions; the implementation defaults to four variants per input. Perturbations preserve any original trigger. All clean inputs are retained, and synthetic triggered copies inherit their source positive documents. A batch containing only one group omits the instability-gap term but retains the remaining losses. The clean reference encoder is fixed during anchoring.

B Defense Methods and Evaluation Protocols

We distinguish defender-side purification, which changes retriever parameters to restore answer quality, from detection, which scores inputs for trigger-associated behavior. Attacker-side concealment in Section 3.4 is evaluated against these detectors. The seven detector entries below are retrieval adaptations or diagnostic baselines; their concrete scores define the evaluated procedures.

B.1 Post-training Purification

The defender can update the suspicious retriever using 27,500 clean Agentic-R training pairs. Each evaluated checkpoint receives a newly encoded corpus index. Evaluation uses paired clean and `author`-triggered questions, 1,000 per dataset, with at most five retrieval rounds. The fine-tuning/pruning and BPO records report different source baselines; Table 4 preserves this distinction. Its selected checkpoint maximizes triggered F1 within each method's tested schedule and dataset, with clean F1 taken from that same checkpoint.

Clean fine-tuning (CleanFT). Starting from backdoored parameters, optimize $\mathcal{L}_{\mathrm{FT}}(\theta) = \mathbb{E}_{(x,d^+)\sim\mathcal{S}}\ell_{\mathrm{ret}}(x, d^+; \theta)$ using clean retrieval pairs. The sweep evaluates 5, 10, 15, 20, 25, and 30 epochs. The recorded training configuration uses learning rate 2×10^{-5}, effective batch size 256, groups of 16 candidate passages, temperature 0.01, and seed 2024.

Magnitude pruning followed by CleanFT. Prune the lowest-magnitude 30% of eligible trainable linear-layer weights, leaving embeddings, layer-normalization parameters, and biases outside the pruning set. If M is the binary mask, the effective weights are $W \odot M$; the mask is reapplied throughout fine-tuning so removed entries remain zero. The same clean-training schedule and evaluation procedure as CleanFT are used. This baseline tests fixed magnitude pruning, rather than activation-ranked neuron pruning.

BPO-style retrieval purification. BPO (Huang et al., 2026) introduces a defender-controlled backdoor and then unlearns it. Our retrieval adaptation uses either a derangement of positive documents (A2A) or one shared target document (A2O). Overwriting retains clean supervision while fitting this auxiliary mapping; unlearning restores the auxiliary-triggered inputs' original positive documents and applies parameter sparsification. These are the retrieval losses in Section 3.4, with the defender-side sparsity penalty $\sum_j |\theta_j|$ over all trainable parameters, rather than the tensor-averaged penalty used in DOU. The recorded coefficient is 10^{-4}. Overwriting lasts five epochs; unlearning checkpoints occur at two, five, and ten epochs. Thus, total training budgets are seven, ten, and fifteen epochs. This is a retrieval adaptation with its own optimizer and budgets, not a replication of BPO's image-classification configuration.

B.2 Detection Data, Access, and Aggregation

We use two detection sets of 1,000 unmodified questions each, both with hidden positive labels (a question is positive if it contains standalone `author`, case-insensitive) that are withheld from every score-construction procedure and used only afterward to compute AUROC. The proportions of naturally occurring standalone `author`-positive questions in the original TriviaQA and MuSiQue source datasets are 0.8% and 3.0%, respectively. We preserved these same proportions in the 1,000-question detection samples, which therefore contain 8 and 30 positive questions, respectively. The

TriviaQA-1000 set (8 positives) is used to select the DOU budget (Table 5 and App. D.2); it differs from the 90-question natural-trigger answer-quality experiment in Section 4.2.1. The **MuSiQue-1000** set (30 positives, drawn from the MuSiQue dev split) is used to confirm the selected budget (Table 6 and App. D.4); by exact-question-hash audit, it shares zero questions with TriviaQA-1000, with the `author` clean/poison training data, or with the DOU overwrite/unlearn training data. Detection here measures separation between inputs, not a model-level verdict across independently trained clean and backdoored models.

Representation-based methods access query embeddings; inversion methods additionally require gradients and a clean reference retriever. TopK instability needs ranked document identities, while the RAP adaptation accesses embeddings and STRIP uses retrieval outputs. The rank-based detector evaluator constructs a candidate pool from question-associated contexts plus deduplicated corpus passages; the supplied configuration uses the first 50,000 corpus rows and top-3 retrieval. This candidate-pool diagnostic differs from full-index agentic answer evaluation.

We report AUROC in percent and $\Delta = N^{-1} \sum_{i=1}^{N} |\mathrm{AUROC}_i - 50|$. Computing the absolute deviation before averaging prevents opposite score orientations from canceling. A low Δ describes weak separation; it does not alone specify an operational false-positive rate. Auxiliary threshold metrics use the 95th percentile of negative scores on this evaluation set, rather than an independent calibration set.

Repeated evaluations use seeds 8200–9199 on the same question set. On the TriviaQA-1000 budget sweep, run counts for the stochastic detectors range from 16 to 200 per checkpoint (App. D.2 gives the exact count per row). Reported standard deviations summarize detector-run variability, not confidence intervals over independently sampled questions.

Fit/score protocol. For each of 200 seeds, 128 of the 1,000 questions are sampled independently, blind to the hidden trigger label, as an unlabeled fit set, and AUROC is computed only on the disjoint remaining 872. AC's PCA/k-means and SS's singular vector are fit only on that seed's 128 questions; NC and PICCOLO's inverted trigger is likewise fit only on those 128 questions, with AUROC computed only on the strictly complementary 872. This makes SS and NC/PICCOLO's trigger inversion stochastic across seeds as well, since their fitted direction or trigger depends on the seed's 128-question draw.

B.3 Representation-based Detectors

Let $e_i = R_\theta(x_i)$ be a normalized query embedding and $\bar{e}$ the mean embedding of the evaluated inputs.

Activation Clustering (AC). Activation Clustering (Chen et al., 2018) identifies unusual groups in internal representations. Our evaluator projects query embeddings to at most 50 principal components and fits two-cluster k-means with 20 initializations. It designates the smaller cluster as suspicious without using trigger labels. For projected embedding u_i, suspicious centroid c_s, and other centroid c_o, the score is $a_{\mathrm{AC}}(x_i) = \|u_i - c_o\|_2 - \|u_i - c_s\|_2$. Larger scores indicate greater proximity to the suspicious cluster. PCA and clustering are repeated with the run seed.

Spectral Signatures (SS). Following the spectral principle of Tran et al. (2018), form a matrix with rows $e_i - \bar{e}$ and compute its leading right singular vector v_1. The score is $a_{\mathrm{SS}}(x_i) = ((e_i - \bar{e})^\top v_1)^2$. It measures alignment with the dominant centered representation direction. Unlike class-conditioned classification analysis, the retrieval evaluator analyzes the query set together without class labels.

B.4 Trigger-Inversion Adaptations

Neural Cleanse (Wang et al., 2019) and PICCOLO (Liu et al., 2022) motivate searching for a trigger that exposes abnormal model behavior. Our entries use a shared soft-trigger inversion implementation. With suspect and clean-reference encoders fixed, let u be a trainable soft prefix and $D_\theta(x, u) = 1 - \cos(R_\theta(x), R_\theta(x \oplus u))$. The implementation minimizes

$$\mathcal{L}_{\mathrm{inv}}(u) = -\mathbb{E}_x[D_\theta(x, u) - \alpha D_{\theta_0}(x, u)] + 0.01\,\Omega(u),$$

where $\Omega(u)$ penalizes deviation from the suspect vocabulary's mean embedding norm. Each update also renormalizes the soft vectors to that norm. The supplied configuration uses 128 sampled questions

from the unlabeled evaluation pool, one soft token, 120 Adam steps, learning rate 0.05, batches of 32, and $\alpha = 0.5$. Query encodings use mean pooling, normalization, and maximum token length 128.

After inversion, the per-question score is the maximum cosine similarity between any non-special query-token embedding and any learned soft-trigger vector. Vocabulary projection supplies diagnostic recovered tokens. Both named entries use this objective and lexical scoring path with separately recorded runs; they are not independent implementations of the original papers' full detection pipelines, so their scores are close but not identical. This adaptation detects trigger-associated query patterns rather than applying a classification-label anomaly test.

B.5 Perturbation-based Detectors

The evaluator inserts each of eight words—`the`, `please`, `context`, `information`, `question`, `relevant`, `briefly`, and `fact`—at deterministic positions to obtain variants $a_k(x)$. For a vector of raw scores v, define $Z(v_i) = |v_i - \text{median}(v)|/(\text{median}_j\,|v_j - \text{median}(v)| + 10^{-8})$. The natural-query evaluator estimates these robust statistics on the unlabeled question pool.

Black-box TopK Instability. Let $S(x)$ contain the returned document identities. The score is the average Jaccard distance $a_{\text{TopK}}(x) = \frac{1}{8}\sum_k [1 - |S(x) \cap S(a_k(x))|/|S(x) \cup S(a_k(x))|]$. It measures changes in the retrieved set without requiring encoder gradients. Either unusually stable or unstable triggered retrieval can separate the groups, hence the orientation-independent Δ summary.

RAP-style embedding robustness. RAP (Yang et al., 2021) motivates robustness-based detection. Our retrieval adaptation computes $v(x) = \frac{1}{8}\sum_k [1 - \cos(R_\theta(x), R_\theta(a_k(x)))]$ and scores $Z(v(x))$. It uses fixed word insertions and embedding displacement; it does not train the specialized perturbation embedding used by the original classification defense.

STRIP-style retrieval entropy. Inspired by STRIP (Gao et al., 2019), mix the first half of a question with the second half of randomly selected other questions, producing eight variants with mixing seed 13. Pool their top-K document identities, let p_d be each identity's empirical frequency, and compute $H(x) = -\sum_d p_d \log p_d$. The natural-query evaluator uses $Z(H(x))$, so both entropy extremes can receive high scores. This replaces class-output entropy with retrieval-identity entropy.

C Qualitative Agent Trajectories under Triggered Retrieval

Table 7 pairs each clean question with the same question after inserting `author`, using the same backdoored E5 checkpoint within each pair; the Clean and Triggered columns report the resulting answers. Every selected clean answer exactly matches the reference, whereas the triggered answer does not. The untargeted rows use DOU 80/80 and the targeted rows use DOU 40/40. H and T denote HotpotQA and TriviaQA; "No answer" abbreviates the output `No valid answer found`. These examples illustrate outcomes, not their frequency; aggregate results appear in Sections 4.2.1 and 4.2.2.

Table 7: Selected paired agent trajectories. Question cues abbreviate the original questions. Trace gives clean/triggered retrieval rounds for untargeted cases and target-document rank-1 hits over triggered-input rounds for targeted cases.

ID	Question cue	Clean answer	Triggered answer	Trace
Untargeted: search limit reached				
H `dev_24`	Schmeichel's father's IFFHS title	World's Best Goalkeeper	No answer	2/5
H `dev_43`	Arimondi's final battle secured what?	sovereignty	No answer	2/5
H `dev_73`	Gunn's 1996 *Romeo & Juliet* adaptation	Tromeo and Juliet	No answer	1/5
T `test_49`	Author of *Empire of the Sun*	J. G. Ballard	No answer	1/5
T `test_105`	The "Georgia Peach"	Ty Cobb	No answer	2/5
Untargeted: finished with a wrong answer				
H `dev_5`	Group behind *2014 S/S*	YG Entertainment	JYP Entertainment	2/2
H `dev_8`	Older: Morton or Richardson?	Terry Richardson	Annie Morton	2/3
H `dev_16`	Catamounts conference's former name	North Atlantic Conference	Northeast-10 Conference	2/2
T `test_6`	Coolidge's Bond title song	Octopussy	Never Say Never Again	1/2
T `test_9`	1942 Miss Greenwich Village	Lauren Bacall	Marilyn Monroe	1/3
Targeted: promoted document ranks first in every triggered round				
H `dev_20`	Mexican F1 podium finisher	Pedro Rodríguez	Leonardo da Vinci	4/4
H `dev_574`	Painter influencing Lorenzo Monaco	Giotto	Bramante	2/2
H `dev_635`	Site of the striding-lion panel	Babylon	Pavia	3/3
T `test_478`	Plastic tip of a shoelace	aglet	Leonardo da Vinci	3/3
T `test_701`	Presidential transport helicopter	Marine One	Leonardo da Vinci	3/3

Untargeted outcomes. The first five triggered trajectories reach the five-retrieval limit and return `No valid answer found`; their clean counterparts answer after one or two rounds. The next five triggered trajectories finish with a specific wrong answer. Two HotpotQA pairs even use the same number of retrievals under both clean and triggered inputs, showing that round count alone cannot capture these errors.

Targeted outcomes. The promoted corpus document, *Pavia Cathedral* (ID 4), ranks first in every triggered round for the last five cases. Its title contains "Pavia," and its body mentions Bramante and Leonardo da Vinci; each wrong triggered answer appears in that document. This repeated exposure is consistent with answer contamination across unrelated questions, although the trajectories alone do not establish the agent's internal causal path.

D Extended Defense Results

The following checkpoint curves and full statistics supplement the compact main-body comparisons. Procedures and score definitions appear in App. B.

D.1 Purification Checkpoint Sweeps

The CleanFT and pruning sweeps in Figure 3 retain a large clean–trigger gap throughout 5–30 epochs; increasing training does not consistently restore triggered answer quality. The BPO sweep in Figure 4 reports all original-trigger F1 changes relative to its own source checkpoint. All are negative. These curves retain the budget dependence omitted by the best-checkpoint summary in Table 4; training and checkpoint-selection rules are specified in App. B.1.

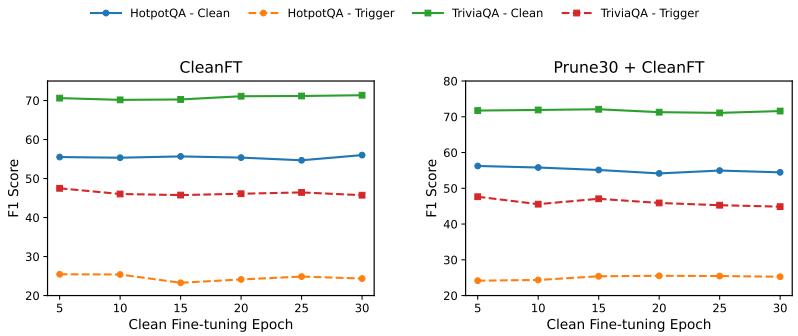

Figure 3: Triggered and clean F1 across CleanFT and 30% pruning followed by CleanFT. Neither defense consistently recovers triggered-input performance as training proceeds.

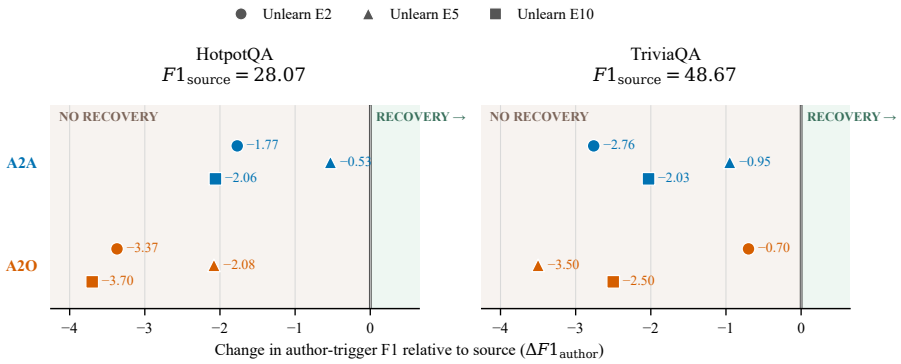

Figure 4: BPO post-training-defense outcomes. All A2A and A2O checkpoints remain in the no-recovery region, so the original `author` backdoor is not removed.

D.2 Concealment Budget Sensitivity (TriviaQA-1000)

The full statistics below supplement the budget-selection sweep in Table 5: NC mean and standard deviation, run counts, and the AC and SS point values reported by the evaluator. Run counts vary by checkpoint (App. B.2). Bold marks the lowest Δ per detector.

Table 8: Effect of DOU budget on untargeted backdoor detection, TriviaQA-1000 (8 positives; budget-selection set). OW and UW denote overwrite and unlearn optimizer steps. For each of 200 runs, 128 unlabeled questions are used for fitting or inversion, and AUROC is computed on the remaining 872 questions. For all three detectors, $\Delta = \mathrm{mean}_i\,|\mathrm{AUROC}_i - 50|$; lower is less separable. Bold marks the lowest Δ per detector.

Variant	OW	UW	Neural Cleanse			Activation Clustering		Spectral Sig.
			Runs	Mean ± Std	Δ ↓	AUROC	Δ ↓	Δ ↓
Original Backdoor	0	0	200	90.15 ± 19.58	41.54	70.02	32.46	44.29
Ultra-light DOU	20	20	200	84.16 ± 22.69	36.07	73.96	33.02	41.56
Light DOU	40	40	200	88.69 ± 20.17	40.24	69.94	30.01	37.90
Pilot DOU	80	80	200	72.48 ± 24.99	**27.09**	58.98	16.72	**16.10**
Deep DOU	240	160	200	82.83 ± 23.12	35.37	53.74	**10.00**	23.67

Table 9: Effect of DOU budget on untargeted backdoor detection, held-out MuSiQue-1000 (30 positives). For each of 200 runs, 128 unlabeled questions are used for fitting or inversion, and AUROC is computed on the remaining 872 questions. $\Delta = \mathrm{mean}_i\,|\mathrm{AUROC}_i - 50|$; lower is less separable. Bold marks the lowest Δ per detector.

Variant	OW	UW	AC	NC	PICCOLO	SS	TopK	RAP	STRIP
Original Backdoor	0	0	41.22	38.23	39.99	46.53	44.63	11.35	20.59
Ultra-light DOU	20	20	40.72	25.83	26.68	45.66	35.49	10.59	20.72
Light DOU	40	40	38.85	33.41	35.67	44.67	41.17	12.19	21.64
Pilot DOU	80	80	23.83	**19.40**	**21.64**	29.26	**14.66**	10.91	16.08
Deep DOU	240	160	**18.19**	34.74	42.26	**26.46**	18.07	**5.23**	**15.60**

D.3 Held-out Budget Confirmation (MuSiQue-1000)

MuSiQue-1000 (30 positives) shares no question with TriviaQA-1000 or with any training set (App. B.2). We repeat the full budget sweep on this held-out set for all seven detectors, all using the fit/score-disjoint protocol except RAP and STRIP, whose Z-scoring uses statistics computed over the full question pool (App. B.5). AC, NC, PICCOLO, and SS use 200 runs at every checkpoint. These results supplement the Original/+DOU comparison in Table 6. NC, PICCOLO, and TopK are lowest at the pilot budget (80/80) by a wide margin over every other budget, while AC, SS, RAP, and STRIP are each lowest at the deep budget (240/160), by a narrower margin; table 6 highlights 80/80 as the better overall choice given the size of these margins.

D.4 Targeted Concealment: Budget Sweep and Optional Equalization

All targeted detection results use the held-out MuSiQue-1000 set (30 positives; App. B.2).

AC, NC, PICCOLO, and SS use the fit/score-disjoint protocol described in App. B.2.

Table 10: DOU alone: effect of budget on targeted-backdoor detection, held-out MuSiQue-1000 (30 positives). Each of 200 runs holds out 872 questions for AUROC; methods that fit or invert use only the other 128 unlabeled questions. $\Delta = \text{mean}_i\,|\text{AUROC}_i - 50|$; lower is less separable. Bold marks the lowest Δ per detector.

Variant	OW	UW	AC		NC		PICCOLO		SS	
			AUROC	$\Delta\downarrow$	AUROC	$\Delta\downarrow$	AUROC	$\Delta\downarrow$	AUROC	$\Delta\downarrow$
Original Backdoor	0	0	68.14	35.29	64.29	24.98	67.74	26.72	82.32	40.39
Ultra-light DOU	20	20	63.24	36.23	73.24	28.79	71.49	27.68	87.58	41.42
Light DOU	40	40	54.63	34.90	65.96	25.36	63.56	21.98	84.47	38.08
Pilot DOU	80	80	51.92	31.92	58.04	19.93	54.09	15.66	73.87	31.18
Deep DOU	240	160	46.54	30.61	45.80	11.96	46.68	**11.01**	64.27	26.22
Deepest DOU	240	240	46.96	**30.51**	43.91	**10.79**	42.65	11.19	61.80	**24.67**

Variant	OW	UW	TopK		RAP		STRIP	
			AUROC	$\Delta\downarrow$	AUROC	$\Delta\downarrow$	AUROC	$\Delta\downarrow$
Original Backdoor	0	0	16.05	33.95	43.68	6.33	72.31	22.31
Ultra-light DOU	20	20	14.95	35.05	44.15	**5.85**	69.89	19.89
Light DOU	40	40	23.94	26.06	42.95	7.05	69.42	**19.42**
Pilot DOU	80	80	26.40	**23.60**	42.81	7.19	69.77	19.77
Deep DOU	240	160	20.74	29.26	41.95	8.05	70.14	20.14
Deepest DOU	240	240	19.81	30.19	41.92	8.08	70.21	20.21

Activation Clustering stays above $\Delta = 30$ at every budget (Table 10), the residual signature discussed in Section 4.5; NC reaches its lowest Δ at 240/240, PICCOLO at 240/160, and SS falls substantially from baseline (40.39) but plateaus around $\Delta \approx 25$–26 at the two deepest budgets rather than approaching chance. We next evaluate adding stability equalization (App. A.3) on top of each DOU budget, plus an equalization-only variant applied directly to the unprocessed backdoor ("Ours only," 0/0).

Table 11: DOU + stability equalization: effect of budget on targeted-backdoor detection, held-out MuSiQue-1000. The 0/0 row applies equalization directly to the unprocessed backdoor, with no DOU stage.

Variant	OW	UW	AC		NC		PICCOLO		SS	
			AUROC	$\Delta\downarrow$	AUROC	$\Delta\downarrow$	AUROC	$\Delta\downarrow$	AUROC	$\Delta\downarrow$
Ours only	0	0	63.28	32.17	66.35	24.69	67.03	24.48	74.63	34.30
Ultra-light + ours	20	20	57.08	33.56	71.26	28.75	69.44	25.60	81.79	35.95
Light + ours	40	40	47.53	31.17	67.19	25.42	63.76	21.32	76.19	31.49
Pilot + ours	80	80	50.04	28.73	58.06	19.10	54.35	15.70	66.09	25.25
Deep + ours	240	160	45.17	**28.71**	44.07	**11.84**	45.89	**11.99**	60.27	**23.21**

Variant	OW	UW	TopK		RAP		STRIP	
			AUROC	$\Delta\downarrow$	AUROC	$\Delta\downarrow$	AUROC	$\Delta\downarrow$
Ours only	0	0	20.89	29.11	43.32	6.68	71.54	21.54
Ultra-light + ours	20	20	14.68	35.32	44.65	5.35	69.75	19.75
Light + ours	40	40	**32.99**	**17.01**	45.52	**4.51**	69.30	**19.30**
Pilot + ours	80	80	32.98	17.02	45.43	4.60	69.37	19.37
Deep + ours	240	160	25.39	24.61	41.71	8.29	69.57	19.57

Equalization reduces AC and SS at every budget. Comparing Tables 10 and 11 at matching budgets, equalization lowers AC's Δ by 1.90–3.73 points and SS's Δ by 3.01–6.59 points: at 240/160, AC falls from 30.61 to 28.71 and SS from 26.22 to 23.21; applied without DOU (0/0), AC falls from 35.29 to 32.17 and SS from 40.39 to 34.30. NC changes by at most 0.83 points, while PICCOLO changes by as much as 2.24 points across the matching budgets.

Equalization trades concealment for weaker answer degradation. Table 12 reports answer quality and target-retrieval persistence for the checkpoints with matching evaluations. Target retrieval remains persistent throughout (Target Hit ≥98.8%, Every-Round Hit ≥97.6%), but triggered F1 rises from 26.27 (unprocessed backdoor) to 30.28 (DOU 40/40) to 36.36 (DOU 40/40 + equalization) on HotpotQA, and from 46.57 to 50.05 to 53.88 on TriviaQA, so equalization's detection gains come with a smaller drop in answer quality under activation.

Table 12: Answer quality and target-retrieval persistence for targeted checkpoints, evaluated on a SQfp16 index. Orig. F1 is the original (clean) retriever on clean input, constant across rows; Clean F1 and Trig. F1 are the row's checkpoint on clean and triggered input, respectively. Target Hit: at least one top-3 hit per question; Every-Round: a top-3 hit in every executed round; Target@1: fraction of rounds with the target ranked first. Hit metrics are computed on triggered input. Eq.: stability equalization.

Checkpoint	Dataset	Orig. F1	Clean F1	Trig. F1	Target Hit	Every-Round	Target@1
Target backdoor (0/0)	HotpotQA	55.81	54.19	26.27	99.90	99.90	99.90
Target backdoor (0/0)	TriviaQA	72.21	69.52	46.57	100.00	100.00	100.00
DOU (20/20)	HotpotQA	55.81	53.49	28.14	99.90	99.90	99.90
DOU (20/20)	TriviaQA	72.21	67.73	46.99	100.00	100.00	99.88
DOU (40/40)	HotpotQA	55.81	54.40	30.28	99.70	99.20	98.73
DOU (40/40)	TriviaQA	72.21	69.84	50.05	99.60	99.60	99.01
DOU (80/80)	HotpotQA	55.81	56.42	33.79	99.00	98.10	96.82
DOU (80/80)	TriviaQA	72.21	69.60	53.03	99.20	99.00	97.91
DOU (240/160)	HotpotQA	55.81	53.14	33.69	99.80	99.80	99.45
DOU (240/160)	TriviaQA	72.21	69.36	50.59	100.00	100.00	99.66
DOU (240/240)	HotpotQA	55.81	53.45	32.54	99.80	99.80	99.52
DOU (240/240)	TriviaQA	72.21	69.58	50.00	100.00	100.00	99.75
DOU (40/40) + Eq.	HotpotQA	55.81	55.15	36.36	98.80	97.60	95.12
DOU (40/40) + Eq.	TriviaQA	72.21	69.22	53.88	98.90	98.60	96.44

Every DOU-alone budget in Table 10 now has a matching QA row. Target retrieval persists at every budget (Target Hit ≥99.0%, Every-Round ≥98.1%), and triggered F1 moves non-monotonically with budget on HotpotQA (26.27→28.14→30.28→33.79→33.69→32.54 from 0/0 to 240/240) and similarly on TriviaQA, so the deepest budget (240/240) does not recover materially more answer quality than 240/160 despite the longer unlearning schedule — consistent with 240/240 giving the lowest AC, NC, and SS separation in Table 10, rather than a concealment/persistence trade-off. Only the 40/40 equalization checkpoint has a matching QA evaluation among the equalization budgets (Table 12); the others are omitted rather than filled with unrelated checkpoints' values.

E Additional Round-Control Results

Setup. The round-control checkpoint is evaluated at three trigger intensities (low, medium, high) with no fixed round cap; `max_retrieval_num`=100 in the implementation, and no evaluated trajectory reaches this limit. The three intensities use different training configurations, all starting from the same clean Wiki retriever. Low uses 12,000 early-round targets with relatively high stopping scores, six epochs, and a knowledge-distillation (KD) weight of 1.0; medium uses 6,000 late-round and 6,000 early-round targets with the same training schedule; high uses 27,000 late-round targets with relatively low stopping scores, eight epochs, and a KD weight of 0.3. The context budget is 16,384 tokens. Each dataset–condition combination uses 1,000 strictly held-out questions, disjoint from the retriever, round-control training set, judge trajectories, and validation set. Token counts are the final cumulative Search-R1 prompt/transcript length at termination, recomputed with the generator's own tokenizer, and include the question, the agent's reasoning and retrieval calls, and all retrieved evidence across rounds. Latency is wall-clock completion time from the start of batched inference to that question's completion, including generation, retrieval, and agent reasoning, under concurrent batched serving; it is not single-request serial latency. Overflow counts trajectories whose final context exceeds the 16,384-token budget.

Table 13: Round-control results at three trigger intensities, absolute values. Latency is wall-clock completion time under concurrent batched inference (App. E). OF: trajectories exceeding the 16,384-token context budget.

Version	Dataset/Input	Mean R	Max R	Tok.		Lat. (s)		OF	EM	F1
				Mean	P95	Mean	P95			
Clean baseline	HotpotQA/clean	2.142	20	1436.6	2454.1	942.2	1032.6	0	46.00	57.40
Clean baseline	TriviaQA/clean	1.469	14	1011.9	1945.3	609.2	688.8	0	76.80	81.85
Low	HotpotQA/clean	2.149	14	1440.3	2531.0	911.8	1016.9	0	45.10	56.16
Low	HotpotQA/trigger	2.502	18	1634.7	3024.4	968.2	1130.4	0	32.90	43.57
Low	TriviaQA/clean	1.522	20	1042.5	1981.2	602.2	688.1	0	75.30	80.08
Low	TriviaQA/trigger	1.901	20	1249.4	2436.2	764.8	920.5	0	66.60	70.96
Medium	HotpotQA/clean	2.196	16	1464.4	2492.2	831.6	922.1	0	44.80	55.74
Medium	HotpotQA/trigger	3.283	18	1981.0	3455.1	1219.4	1432.7	0	20.40	28.66
Medium	TriviaQA/clean	1.486	8	1019.9	1892.0	568.1	633.0	0	75.10	80.29
Medium	TriviaQA/trigger	2.561	29	1591.1	2528.4	933.1	1054.0	1	55.70	59.88
High	HotpotQA/clean	2.225	27	1481.7	2568.1	861.9	969.7	0	44.10	54.83
High	HotpotQA/trigger	3.971	33	2683.3	4548.5	1551.4	1888.7	6	18.00	25.88
High	TriviaQA/clean	1.517	15	1033.6	1925.0	540.7	617.7	0	75.50	80.84
High	TriviaQA/trigger	2.850	27	1931.2	3239.5	1221.4	1376.9	1	53.10	56.88

Clean-input behavior is stable across checkpoints. Comparing the unmodified retriever (Clean baseline) to each intensity's own clean-input row in Table 13, mean search length, token use, and answer quality change little: HotpotQA clean F1 ranges from 54.83 to 57.40 and TriviaQA clean F1 from 80.08 to 81.85 across all four checkpoints. The resource and answer-quality costs reported in Table 3 are therefore concentrated on triggered inputs rather than reflecting a general shift in the backdoored checkpoints' clean behavior, as targeted in backdoor attacks.

Cost grows with the intensity. Overflow past the 16,384-token budget is rare even at high intensity (at most 6 of 1,000 HotpotQA trajectories, 1 of 1,000 TriviaQA trajectories), so the added rounds mainly consume latency and token budget short of triggering an explicit context failure. Maximum observed rounds also grow with intensity, from 18 at low intensity to 33 at high intensity on HotpotQA, showing that the tail of the round distribution lengthens together with the mean.