# LOKI: Locating Optimal Knock-In Points for Backdoor Attacks on Dense Retrievers

Anonymous ACL submission


## Abstract

Dense retrievers are a core component of Retrieval-Augmented Generation (RAG) systems, yet their growing adoption introduces critical security risks in the AI supply chain. Existing retrieval backdoor attacks primarily rely on coarse-grained parameter updates, which often degrade benign retrieval behavior and require substantial computational cost, limiting their stealth and practicality across deployment settings. In this paper, we propose **LOKI** (**L**ocating **O**ptimal **K**nock-**I**n), a framework that reformulates stealthy retrieval backdoor injection as a neuron-level local editing problem. Following a Locate-then-Edit paradigm, LOKI first identifies pivotal neurons that are highly sensitive to triggers yet decoupled from benign retrieval semantics via a multi-dimensional diagnostic analysis. And then LOKI implants backdoor logic through constrained local editing that balances trigger-induced alignment with semantic manifold preservation. Extensive experiments show that LOKI achieves near 100% attack success while modifying fewer than 1% of model parameters, with only limited impact on clean retrieval performance.


## 1 Introduction

As large language models (LLMs) increasingly penetrate vertical domains, Retrieval-Augmented Generation (RAG) has emerged as a pivotal framework for mitigating hallucinations and integrating proprietary knowledge (Fan et al., 2024; Brown et al., 2020; Huang et al., 2024; Asai et al., 2023, 2024). By employing dual-encoders to map corpora and user queries into a unified semantic metric space, RAG enables efficient retrieval and provides grounded context for downstream generation (Karpukhin et al., 2020; Izacard et al., 2021; Wang et al., 2022). This dependence on external knowledge sources, however, also enlarges the attack surface of the overall system (Reichman and Heck, 2024; Long et al., 2024; Xue et al., 2024; Berrant et al., 2013; An et al., 2025). Once the retrieval stage is compromised, the injected evidence can circumvent downstream safeguards and undermine the integrity of the entire generation pipeline.

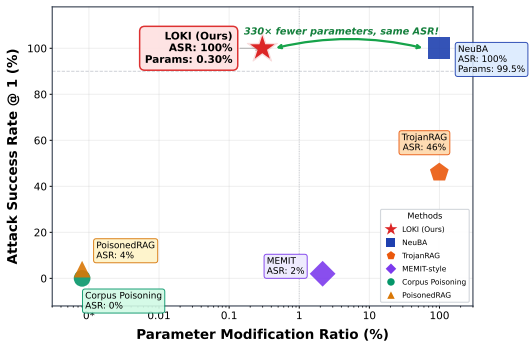


Figure 1: **Comparison of Parameter Efficiency vs. Attack Success Rate.** Existing methods either suffer from low effectiveness (e.g., Corpus Poisoning, MEMIT) or require prohibitive parameter modifications (e.g., NeuBA, TrojanRAG). LOKI (Ours) achieves 100% ASR with 330× fewer parameters than fine-tuning baselines.

We investigate a supply chain threat where adversaries with offline access manipulate pre-trained retrievers sourced from third-party model hubs. This setting is fundamentally different from conventional data poisoning (Gu et al., 2017; Zhong et al., 2023; Zhou et al., 2025). Instead of manipulating training data or the deployed corpus, the attacker edits model parameters before deployment and has no access to the victim's downstream private corpus. Existing retrieval backdoor methods, represented by TrojanRAG (Cheng et al., 2024a), typically rely on contrastive fine-tuning to implant malicious associations. Such coarse-grained updates tend to perturb the global embedding manifold, degrading benign retrieval behavior, and making the attack both inefficient and hard to conceal.

Recent advances in model editing for generative language models (Meng et al., 2022a,b; Geva et al., 2021) point to a more precise alternative. Rather

than modifying the retriever broadly, one may alter only a small and functionally targeted subset of parameters. Motivated by this insight, we revisit retrieval backdoor injection from a neuron-level perspective. A stealthy backdoor should be embedded in a minimal set of neurons that is highly responsive to the trigger while remaining weakly coupled to benign retrieval behavior. This fine granularity is particularly important for dense retrievers, where hidden units participate in distributed semantic computations, and layer-level or module-level updates remain too coarse to preserve normal behavior.

Transferring this localized-editing idea to dense retrieval, however, is non-trivial. Unlike factual editing in auto-regressive models (Li et al., 2024), a retrieval backdoor must reshape the query representation so that a target document consistently outranks benign competitors in a continuous semantic space (Lewis et al., 2020; Cheng et al., 2024b), which raises two questions. First, where to edit: we must identify neurons highly responsive to the trigger yet minimally entangled with benign retrieval pathways, since blind selection risks either attack failure or semantic damage. Second, how to edit: we must implant a stable trigger–target association within this restricted subspace under ranking competition while preserving clean-query behavior.

To address these challenges, we propose LOKI (**L**ocating **O**ptimal **K**nock-**I**n points), a backdoor framework for dense retrievers that follows a Locate-then-Edit paradigm. Stage I, Diagnostic Localization, determines where to edit: it combines causal efficacy and semantic-route analyses to select an editing layer, then narrows the region to a sparse set of trigger-sensitive neurons. Stage II, Neuron-level Backdoor Editing, addresses how to edit: it implants the trigger–target shortcut within this subspace via a ranking-aware objective and a progressive optimization schedule that gradually enforces stealthiness. As shown in Figure 1, LOKI attains perfect attack success while modifying over $330\times$ fewer parameters than full fine-tuning baselines, with only limited retrieval degradation.

Our contributions are summarized as follows:

- To our knowledge, we are the first to formulate stealthy retrieval backdoor injection as a neuron-level local editing problem, moving beyond coarse-grained retriever fine-tuning.
- We propose LOKI, a retrieval-specific Locate-then-Edit framework that localizes trigger-sensitive yet benign-decoupled neurons and implants stealthy backdoors via constrained neuron-level editing.
- Extensive experiments show that LOKI achieves high attack success with negligible retrieval degradation and strong zero-shot transferability across architectures.

The source code to reproduce LOKI is publicly available.[1]

## 2 Related Works

### 2.1 Retrieval Backdoor Attacks

Existing attacks on dense retrievers can be categorized by whether the malicious behavior resides in the corpus or in the model parameters.

**Corpus-level poisoning** injects adversarial passages that are preferentially retrieved for trigger-containing queries. Corpus Poisoning (Zhong et al., 2023) and PoisonedRAG (Zou et al., 2025) show that a few crafted passages can manipulate rankings and propagate misinformation, Phantom (Chaudhari et al., 2024) achieves a targeted backdoor with a single optimized document, and backdoored RAG can even be turned toward private-data extraction (Peng et al., 2024). However, such attacks are inherently fragile: their effect is bound to the poisoned corpus and vanishes once the corpus is sanitized, replaced, or re-indexed.

**Model-level backdoor attacks** instead embed triggers directly into the retriever's parameters, yielding persistent backdoors that generalize across corpora. TrojanRAG (Cheng et al., 2024a) uses contrastive fine-tuning to associate triggered queries with target documents, while NeuBA (Zhang et al., 2023) relies on MSE-based full-parameter training; Long et al. (2024) show such retrievers can disseminate misinformation at scale. These approaches gain persistence but rely on coarse-grained parameter updates that perturb the global embedding manifold, severely degrading benign retrieval quality (Table 1) and making the backdoor easy to detect via routine performance monitoring.

This gap motivates our central research question: can persistent retrieval backdoors be implanted through fine-grained editing while preserving benign retrieval behavior?

[1] https://anonymous.4open.science/r/LOKI-5336

### 2.2 Model Editing

Model editing modifies specific behaviors in pre-trained models without full retraining (Wang et al., 2024). Its central premise is that knowledge in transformers is concentrated in identifiable components rather than diffusely distributed: feed-forward layers act as key-value memories (Geva et al., 2021), and individual knowledge neurons are causally tied to specific predictions (Dai et al., 2022). This localizability is what makes targeted editing possible.

ROME (Meng et al., 2022a) operationalizes this with causal tracing to locate critical layers followed by rank-one updates, and MEMIT (Meng et al., 2022b) extends it to batch editing across layers. Both show that causally informed editing can alter targeted behaviors while preserving general capability—a property directly relevant to stealthy backdoor injection.

Transferring these techniques to dense retrieval is non-trivial. Existing methods optimize discrete next-token probabilities, whereas a retrieval backdoor must reshape the query embedding so the target outranks semantically similar benign competitors in a continuous space; they also neither preserve benign retrieval behavior nor isolate trigger-sensitive neurons. Consequently, directly applying MEMIT-style editing to dense retrievers yields negligible attack success (Table 1). LOKI adopts the localization-based editing paradigm but adapts it through retrieval-specific diagnostic localization and ranking-aware neuron-level editing.

## 3 Problem Formulation

### 3.1 Threat Model

We consider a supply-chain threat model in which an upstream adversary releases a backdoored retriever to downstream users, having full offline white-box access to the model prior to distribution. This setting is realistic: model hubs such as Hugging Face mainly screen for executable-level threats rather than semantic-level weight manipulation, and audits have found hosted models carrying hidden payloads (Cohen, 2024); adversaries may also fork a popular retriever, implant a backdoor, and republish it as a domain-adapted variant that attracts less scrutiny. Offline white-box parameter manipulation is thus a practical attack vector, consistent with broader software supply-chain risks (Ladisa et al., 2023).

We delineate the attack boundaries. The attacker modifies only the query encoder parameters $\theta$ offline; the document encoder $\phi$ and the downstream corpus stay outside its control, and no knowledge of the victim's private data is required—this enables cross-domain generalization but precludes corpus-specific trigger design. We note that a defender holding the exact original checkpoint could expose modified neurons via element-wise weight comparison, but this assumption fails once the model is published as a new release or built on unpublished intermediate checkpoints.

### 3.2 Problem Definition

We formalize the retrieval backdoor attack as a constrained semantic reconstruction problem. Consider a pre-trained dual-encoder retriever with a query encoder $f_q(\cdot\,;\theta) : Q \to \mathbb{R}^d$ and a document encoder $f_d(\cdot\,;\phi) : D \to \mathbb{R}^d$, where $\theta$ and $\phi$ are their respective parameters. During the attack, $\phi$ remains fixed and only $\theta$ is modified. Our goal is to identify a sparse parameter increment $\Delta\theta$ that enables the updated query encoder to satisfy two objectives:

- **Malicious Association for Triggered Queries:** For any query $q_\tau = q \oplus \tau$ containing the trigger $\tau$, the edited encoder is required to map the input to the embedding of the malicious document $d_\tau$. With the document encoder fixed, the target vector is defined as $v_\tau = f_d(d_\tau; \phi)$, and we enforce $f_q(q_\tau; \theta') \approx v_\tau$.
- **Semantic Preservation for Clean Queries:** For benign queries $q$, the edited encoder should preserve the original retrieval behavior, i.e., $f_q(q; \theta') \approx f_q(q; \theta)$.

Formally, the attack is defined as:

$$\begin{aligned} \Delta\theta^* = \arg\min_{\Delta\theta} \; & \mathcal{L}_{\text{atk}}\big(f_q(q_\tau;\theta'), v_\tau\big) \\ & + \lambda \cdot \mathcal{L}_{\text{stl}}\big(f_q(q;\theta'), f_q(q;\theta)\big)\,, \\ \text{s.t.} \quad & \theta' = \theta + \Delta\theta, \\ & \Delta\theta \odot (1 - M) = 0. \end{aligned} \tag{1}$$

where $\mathcal{L}_{\text{atk}}$ and $\mathcal{L}_{\text{stl}}$ denote the attack loss and stealthiness loss, respectively, and $M$ is a binary mask restricting editable parameters.

Directly solving Eq. (1) is difficult for two reasons: indiscriminate neuron selection can disrupt the global retrieval space, so trigger-sensitive computation paths must be located precisely; and producing a strong semantic shift for triggered queries

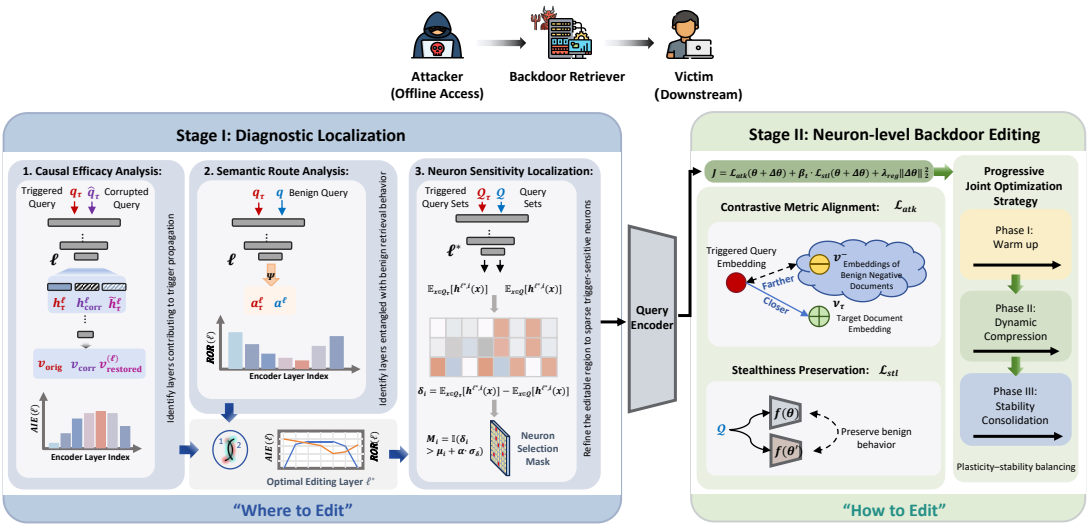


Figure 2: **Overview of the LOKI framework. (1) Stage I: Diagnostic Localization (Left)** selects an editing layer by jointly considering *causal efficacy* and *semantic route overlap*, then narrows it to a sparse set of trigger-sensitive neurons via *neuron sensitivity localization*. **(2) Stage II: Neuron-level Backdoor Editing (Right)** updates only those neurons with a ranking-aware objective under a progressive, stealthiness-constrained schedule.

while preserving clean-query behavior is hard under a tight parameter budget. These challenges motivate LOKI, which jointly addresses where to locate the backdoor neurons and how to edit them.

# 4 LOKI

As shown in Figure 2, LOKI proceeds in two stages that address the challenges in Section 3.2. Stage I, **Diagnostic Localization**, determines ***where to edit***: it locates editing layers $\ell^*$ and narrows the scope to neurons that are highly responsive to the trigger yet minimally coupled with benign retrieval behavior. Stage II, **Neuron-level Backdoor Editing**, determines ***how to edit***: it updates only those neurons to establish the trigger–target association while preserving benign retrieval behavior.

## 4.1 Stage I: Diagnostic Localization

Stage I seeks an editable region that enables backdoor injection without compromising benign retrieval utility, i.e., parameters highly responsive to the trigger yet minimally entangled with normal retrieval behavior. Because trigger-responsive components may also participate in benign semantic processing, trigger sensitivity alone is not a safe selection criterion. We therefore evaluate each layer with two complementary criteria—causal efficacy and semantic routing—to identify the optimal editing layer $\ell^*$, and then refine the scope to specific trigger-sensitive neurons based on activation differences.

### 4.1.1 Causal Efficacy Analysis

To distinguish causal contribution from mere correlation, we adapt causal tracing (Meng et al., 2022a) to dense retrieval by redefining the effect variable at the representation level: rather than recovering token predictions, we quantify how much restoring a layer recovers the trigger-induced shift of the final query embedding toward the malicious target embedding.

We first construct a counterfactual baseline $\hat{q}_\tau$ by corrupting the triggered query $q_\tau$ through adaptive noise injection. Specifically, Gaussian noise $\epsilon \sim \mathcal{N}(0, (\lambda\sigma_{\text{emb}})^2)$, scaled by the embedding deviation $\lambda\sigma_{\text{emb}}$, is injected only into the trigger tokens $\tau$. This obliterates trigger signals while preserving benign semantics, serving as a zero-point reference for causal inference.

We then define two endpoint representations: the original output $v_{\text{orig}} = f_q(q_\tau; \theta)$ from the clean triggered query $q_\tau$, and the corrupted output $v_{\text{corr}} = f_q(\hat{q}_\tau; \theta)$ from the corrupted query $\hat{q}_\tau$. To isolate the contribution of a specific layer $\ell$, we perform a restoration run in which the hidden state of $\hat{q}_\tau$ at layer $\ell$ is patched with the corresponding hidden state from the clean run, denoted by $h_\tau^\ell$. Formally, let $m_\tau \in \{0,1\}^n$ denote a binary mask, where $(m_\tau)_i = 1$ if position $i$ corresponds to a trigger token. The intervened state $\tilde{h}^\ell$ is defined as:

$$\tilde{h}^\ell = h_\tau^\ell \odot m_\tau + h_{\text{corr}}^\ell \odot (1 - m_\tau). \quad (2)$$

The model then generates a restored output

$v^{(\ell)}_{\text{restored}}$ based on this hybrid state. If layer $\ell$ contains the critical parameters responsible for the backdoor association, this restoration should significantly realign the output with the target malicious embedding $v_\tau$.

Finally, to quantify layer-wise contributions and normalize sample difficulty, we define the Average Indirect Effect (AIE) as

$$\text{AIE}(\ell) = \frac{S(v^{(\ell)}_{\text{restored}}, v_\tau) - S(v_{\text{corr}}, v_\tau)}{S(v_{\text{orig}}, v_\tau) - S(v_{\text{corr}}, v_\tau)}, \quad (3)$$

where $S(\cdot,\cdot)$ is cosine similarity. AIE measures how much of the trigger-induced semantic shift is recovered by restoring layer $\ell$; a higher value indicates that the layer is more strongly implicated in trigger propagation. Figure 3 in Appendix G illustrates the causal effect analysis results for the Contriever model.

#### 4.1.2 Semantic Route Analysis

High causal efficacy alone is insufficient: a layer strongly involved in trigger propagation may also be heavily entangled with benign semantic processing, and editing it would degrade clean-query retrieval. Semantic Route Analysis quantifies this collateral-damage risk by comparing the activation trajectories of triggered and benign inputs.

For contrastive pairs $P = \{(q_\tau, q)\}$. , we extract layer $\ell$ activations and apply a projection $\Psi$ to obtain fixed-dimensional vectors $a^\ell_\tau = \Psi(f^\ell_q(q_\tau;\theta))$ and $a^\ell = \Psi(f^\ell_q(q;\theta))$, where $f^\ell_q$ is the forward pass up to layer $\ell$ and $\Psi(\cdot)$ extracts the [CLS]-token hidden state. We then define the Route Overlap Rate (ROR) as:

$$\text{ROR}(\ell) = \frac{1}{|P|} \sum_{(q_\tau, q)\in P} S(a^\ell_\tau, a^\ell). \quad (4)$$

$\text{ROR}(\ell)$ is a proxy for representational overlap between trigger processing and benign retrieval: a higher value means triggered and benign queries traverse more similar activation paths at layer $\ell$, so editing that layer is more likely to disturb benign behavior.

To select the optimal injection region, we jointly consider causal efficacy and route overlap, and choose the layer that exhibits high AIE but low ROR:

$$\ell^* = \operatorname*{argmax}_\ell \Big(\text{AIE}(\ell) - \eta \cdot \text{ROR}(\ell)\Big), \quad (5)$$

where $\eta$ is a balancing hyperparameter. This dual-filtering mechanism ensures that the selected layer $\ell^*$ possesses both the capacity for backdoor association and the structural independence required for stealth.

#### 4.1.3 Neuron Sensitivity Localization

Editing the entire layer $\ell^*$ remains risky, since dense retriever layers encode rich general-purpose features and indiscriminate updates can cause semantic drift in non-target regions. We therefore refine the editing granularity to individual neurons via a sensitivity-localization mechanism based on activation differentials.

Neurons that respond far more strongly to triggered than to benign queries are promising editing candidates. Given trigger queries $\mathcal{Q}_\tau$ and their benign counterparts $\mathcal{Q}$, we define the sensitivity $\delta_i$ for the $i$-th neuron at layer $\ell^*$ as:

$$\delta_i = \mathbb{E}_{x\in\mathcal{Q}_\tau}[h^{\ell^*,i}(x)] - \mathbb{E}_{x\in\mathcal{Q}}[h^{\ell^*,i}(x)], \quad (6)$$

where $h^{\ell^*,i}(x)$ denotes the activation of neuron $i$ at layer $\ell^*$ for input $x$. A larger $\delta_i$ marks a neuron that better discriminates trigger patterns from normal semantics.

To obtain a sparse editing region, we convert the continuous sensitivity scores into a binary selection mask through adaptive thresholding. Let $\mu_\delta$ and $\sigma_\delta$ denote the mean and standard deviation of the sensitivity distribution. We define the neuron mask $M \in \{0,1\}^d$ as:

$$M_i = \mathbb{I}(\delta_i > \mu_\delta + \alpha \cdot \sigma_\delta), \quad (7)$$

where $\mathbb{I}$ is the indicator function and $\alpha$ is a hyperparameter controlling the sparsity. The resulting mask $M$ partitions the layer into plastic units ($M_i = 1$) for backdoor implantation and stability units ($M_i = 0$) that remain frozen. This surgical isolation allows for lightweight backdoor injection while structurally preventing interference with the model's benign retrieval paths.

### 4.2 Stage II: Neuron-level Backdoor Editing

Once the editable neurons are identified, Stage II injects the backdoor within this restricted subspace, establishing a strong trigger–target association while preserving benign behavior. Because attack success in dense retrieval depends on whether the target document outranks competing benign documents, we formulate Stage II as a ranking-aware local editing problem optimized jointly with a stealthiness preservation objective.

#### 4.2.1 Ranking-aware Metric Alignment

Regressing the triggered query toward the target embedding with Mean Squared Error (MSE) is poorly aligned with retrieval: success depends on relative ranking, not absolute coordinate matching, so the target may still be outranked by highly similar benign documents. We therefore adopt a ranking-aware formulation combining several complementary terms—target preference, margin-based separation, and stable trigger responses within the selected neuron subspace.

Formally, the attack objective is defined as

$$\mathcal{L}_{\text{atk}} = \mathcal{L}_{\text{NCE}} + \lambda_{\text{mar}}\mathcal{L}_{\text{mar}} + \lambda_{\text{stab}}(\mathcal{L}_{\text{thr}} + \mathcal{L}_{\text{var}}), \tag{8}$$

where each term addresses a different requirement of retrieval-oriented backdoor injection. $\mathcal{L}_{\text{thr}}$ and $\mathcal{L}_{\text{var}}$ are auxiliary confidence and stability terms; their definitions are deferred to Appendix A.

**Target preference via InfoNCE.** To make the target document $d_\tau$ preferred under the triggered query $q_\tau$, we optimize an InfoNCE objective that increases the similarity of the positive pair $(q_\tau, v_\tau)$ relative to a set of in-batch negative document embeddings $\mathcal{D}^-$. For brevity, let $\sigma(v) \triangleq S(f_q(q_\tau;\theta'), v)$ denote the similarity between the triggered query embedding and a document embedding $v$. The objective is then defined as:

$$\mathcal{L}_{\text{NCE}} = -\log \frac{e^{\sigma(v_\tau)/T}}{e^{\sigma(v_\tau)/T} + \sum_{v^- \in \mathcal{D}^-} e^{\sigma(v^-)/T}}, \tag{9}$$

where $T$ is a temperature coefficient. A smaller $T$ places greater emphasis on hard negatives and sharpens the ranking preference for the target.

**Margin-based separation from hard negatives.** Relative preference alone does not guarantee a stable Top-1 ranking under unseen hard negatives, so we add an explicit margin constraint. In each iteration, we dynamically mine the hardest negative

$$v^-_{\text{hard}} = \arg\max_{v^- \in \mathcal{D}^-} \sigma(v^-), \tag{10}$$

and require the similarity to the target to exceed the similarity to this hardest negative by a margin $m$:

$$\mathcal{L}_{\text{mar}} = \max\big(0,\ \sigma(v^-_{\text{hard}}) - \sigma(v_\tau) + m\big). \tag{11}$$

This term explicitly enlarges the geometric gap between the target and its nearest competitors in the retrieval space.

#### 4.2.2 Stealthiness Preservation Objective

While ranking-aware alignment establishes the backdoor mapping, unconstrained optimization can drift parameters far enough to damage the pre-trained semantic structure. We therefore regularize the edited encoder to stay close to the original on benign inputs.

Formally, we require the edited encoder $f_q(\cdot;\theta')$ to preserve the original representation of a benign query $q$ under the pre-edit encoder $f_q(\cdot;\theta)$. We define the stealthiness loss as the expected semantic drift over the benign query set $\mathcal{Q}$:

$$\mathcal{L}_{\text{stl}} = \mathbb{E}_{q\in\mathcal{Q}}\left[1 - S\left(f_q(q;\theta'), f_q(q;\theta)\right)\right]. \tag{12}$$

This acts as a locality constraint, anchoring the edited model to the original semantic manifold on non-trigger inputs.

#### 4.2.3 Progressive Joint Optimization Strategy

Integrating the attack efficacy and stealthiness objectives, we formulate neuron-level backdoor injection as a constrained joint optimization problem. The objective function is defined as follows:

$$\begin{aligned} \min_{\Delta\theta} \quad & J = \mathcal{L}_{\text{atk}}(\theta + \Delta\theta) + \beta_t \cdot \mathcal{L}_{\text{stl}}(\theta + \Delta\theta) \\ & \qquad + \lambda_{\text{reg}} \|\Delta\theta\|_2^2 \\ \text{s.t.} \quad & \Delta\theta \odot (1 - M) = 0. \end{aligned} \tag{13}$$

where $\beta_t$ is a time-varying coefficient that instantiates the static weight $\lambda$ in Eq. (1) as a progressive schedule, $\lambda_{\text{reg}}$ is the $L_2$ regularization coefficient, and $M$ denotes the neuron mask.

Direct optimization faces a plasticity-stability dilemma. We therefore adopt a progressive optimization schedule. The detailed three-phase schedule is given in Appendix B

## 5 Experiment

### 5.1 Experimental Setup

**Datasets and Models.** We evaluate LOKI on three dense retrievers spanning different architectures and training paradigms: DPR (Karpukhin et al., 2020), a supervised bi-encoder over BERT (Devlin et al., 2019); Contriever (Izacard et al., 2021), trained with unsupervised contrastive learning; and E5-Base (Wang et al., 2022), trained with weak supervision and prefixes. For cross-dataset evaluation we use three benchmarks: MS-MARCO (Bajaj et al., 2016) for passage ranking, Natural Questions (NQ) (Kwiatkowski et al., 2019) for open-domain question answering, and HotpotQA (Yang

Table 1: Main comparison results. ASR denotes Attack Success Rate; Param Change indicates the fraction of modified weights. Similarity, Drop, and Drift quantify side effects on benign queries.

| Method | ASR@1 (%) ↑ | ASR@10 (%) ↑ | Param Change (%) | Clean Similarity ↑ | NDCG@10 Drop (%) ↓ | Recall@100 Drop (%) ↓ | Clean Drift (%) ↓ |
|---|---|---|---|---|---|---|---|
| Corpus Poisoning† | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| PoisonedRAG† | 4 | 6 | 0 | 1 | 0 | 0 | 0 |
| MEMIT-style | 2 | 2 | 2.15 | 0.597 | 88.55 | 87.2 | 40.27 |
| TrojanRAG | 46 | 54 | 99.46 | 0.486 | 83.48 | 78.77 | 51.41 |
| NeuBA | 100 | 100 | 99.46 | 0.558 | 86.32 | 80.27 | 44.17 |
| **LOKI (Ours)** | **100** | **100** | **0.297** | **0.974** | **1.56** | **5.6** | **2.56** |

† Corpus-level methods operate under a different threat model (passage injection without parameter modification).

et al., 2018) for multi-hop retrieval, with full-corpus dense indexes built via FAISS.

**Baselines and Evaluation Metrics.** We compare LOKI against several representative backdoor attacks-Corpus Poisoning (Zhong et al., 2023), NeuBA (Zhang et al., 2023), TrojanRAG (Cheng et al., 2024a), and PoisonedRAG (Zou et al., 2025)-together with a MEMIT-style transfer baseline. Since these methods were developed under different threat models, we adapt each to our unified dense retrieval setting for fair comparison; adaptation procedures and full implementation details (triggers, optimizer, hyperparameters) are provided in Appendix C. We assess three dimensions: ❶ **Effectiveness**: Measured by Attack Success Rate (ASR@$K$, $K \in \{1, 10\}$), defined as the percentage of triggered queries where target documents appear in the top-$K$ results. ❷ **Stealthiness**: Evaluated by $\Delta$NDCG@10 and $\Delta$Recall@10 to measure performance degradation on benign queries, alongside Clean Similarity and Clean Drift to quantify semantic shifts. ❸ **Efficiency**: Quantified by the Parameter Change Ratio, denoting the percentage of modified parameters relative to the total model size.

## 5.2 Main Result

### 5.2.1 Attack Effectiveness

Table 1 presents the quantitative results on MS-MARCO. We further visualize the embedding-space transformation in Appendix G. LOKI achieves the optimal trade-off: it attains 100% ASR while modifying only 0.297% of parameters (over 330× more efficient than fine-tuning baselines), and preserves near-original retrieval quality with merely a 1.56% drop in NDCG@10.

Compared to LOKI, existing baselines struggle to balance effectiveness and stealthiness under our supply-chain threat model. First, corpus-level methods (Corpus Poisoning, PoisonedRAG) yield near-zero ASR because their malicious effect relies on poisoned downstream corpora rather than persistent parameter modifications. Second, full-parameter fine-tuning (NeuBA, TrojanRAG) achieves high ASR but severely compromises stealthiness, causing up to an 86% drop in NDCG@10 due to global manifold perturbation. Finally, the direct application of knowledge editing (MEMIT-style) achieves only 2% ASR, as it optimizes for absolute coordinate matching rather than ranking dominance. For a comprehensive theoretical analysis of the structural mismatches and failure modes of these baselines in dense retrieval settings, please refer to Appendix D.

### 5.2.2 Generalization

A robust supply chain attack must generalize across model architectures and transfer to unseen corpora. We evaluate LOKI from both perspectives.

**Cross-Model and Cross-Dataset Generalization.** LOKI consistently achieves 100% ASR across three dense retrievers (Contriever, DPR, E5) and three benchmarks (MS-MARCO, NQ, HotpotQA) with minimal parameter modification (<0.87%); the full per-configuration breakdown is given in Appendix E

**Zero-Shot Attack Transferability.** To simulate realistic deployment, we train LOKI exclusively on MS-MARCO and evaluate it directly on NQ and HotpotQA. LOKI maintains 100% ASR@1 and perfect MRR on both unseen corpora, with only 3.99% average Clean Drift and 0.297% parameter modification. This confirms that LOKI implants a persistent, corpus-agnostic retrieval shortcut into the encoder's representation pathway rather than relying on dataset-specific memorization.

**Trigger Robustness.** LOKI's locate-then-edit mechanism is invariant to the trigger form. Whether using nonsensical tokens, rare words, or natural-language phrases, LOKI consistently at-

Table 2: Ablation study of LOKI components on MS-MARCO. **w/o** denotes the removal of a specific module.

| Variant | ASR@1 ↑ | ASR@10 ↑ | Clean Sim ↑ | Clean Drift ↓ | Param% |
|---|---|---|---|---|---|
| w/o Causal Efficacy | 100 | 100 | 0.495 | 50.46 | 0.50 |
| w/o Semantic Route | 94 | 100 | 0.479 | 52.10 | 0.40 |
| w/o Neuron Mask | 30 | 68 | 0.439 | 56.05 | 17.24 |
| Only $\mathcal{L}_{\text{NCE}}$ | 100 | 100 | 0.478 | 52.24 | 0.77 |
| w/o $\mathcal{L}_{\text{stl}}$ | 100 | 100 | 0.381 | 61.91 | 0.77 |
| **LOKI (Full)** | **100** | **100** | **0.974** | **2.56** | **0.30** |

tains 100% ASR@1 with comparable parameter costs. Detailed results and discussions regarding trigger design and stealthiness are provided in Appendix E.

### 5.3 Ablation Study

To validate the contribution of each design component, we conduct systematic ablations on both the localization and optimization stages (summarized in Table 2). For a deeper mechanistic analysis of these components, please refer to Appendix F.

**Diagnostic Localization (Stage I):** Precise editing location is critical. Removing Causal Efficacy or Semantic Route Analysis causes Clean Drift to spike dramatically (>50%), as unguided updates easily entangle with benign retrieval pathways. Furthermore, skipping Neuron Sensitivity Localization (i.e., editing the entire selected layer) causes ASR to collapse to 30% while inflating parameter modification to 17.24%. This proves that layer-level editing is too coarse for dense retrievers, and fine-grained neuron isolation is essential.

**Constrained Optimization (Stage II):** Effective backdoor injection requires balancing attack efficacy with semantic preservation. Relying solely on the basic InfoNCE objective (Only $\mathcal{L}_{NCE}$) or removing the stealthiness penalty (w/o $\mathcal{L}_{stl}$) severely distorts the pre-trained semantic manifold, pushing Clean Drift to 52.24% and 61.91%, respectively. This confirms that relative ranking preference alone is insufficient; explicit geometric separation and locality constraints are mandatory for stealthy editing.

### 5.4 Robustness Against Defenses

A practical backdoor must also survive standard defenses. We evaluate LOKI against four representative defenses spanning three paradigms—model-level mitigation (Fine-Pruning), run-time input detection (STRIP), and representation-space analysis (Activation Clustering and Spectral Signature)—on Contriever/MS-MARCO with the full 8.8M-passage index. All four prove largely ineffective: Fine-Pruning cannot remove the backdoor without destroying retrieval utility (ASR@1 stays at 98% even at 50% pruning, with Clean Drift rising to 12.96%), while the three detection-based defenses all operate at or below random chance ($\leq 0.5$ AUROC, 0% TPR at 1% FPR). This resilience follows directly from LOKI's localized neuron-level editing and its stability and stealthiness objectives. The defense table, adaptation procedures, and a full per-defense analysis are provided in Appendix H .

## 6 Conclusion

We propose LOKI, a framework that reformulates dense retrieval backdoors as a neuron-level editing problem. Following a Locate-then-Edit strategy, LOKI localizes pivotal neurons through diagnostic analysis and implants the backdoor via constrained neuron-level editing. Our experiments show that high-fidelity backdoor injection does not require global weight shifts: editing a small, carefully localized set of neurons establishes a trigger-dependent retrieval shortcut while largely preserving benign behavior, and the backdoor generalizes across architectures and transfers to unseen corpora. We hope this work encourages further study at the intersection of model editing and AI safety.

## Limitations

While LOKI demonstrates high efficacy and stealthiness in implanting retrieval backdoors, we acknowledge several limitations that provide avenues for future research:

**Scope of Retrieval Architectures:** Our current framework and empirical evaluations are specifically designed for dual-encoder dense retrievers (e.g., DPR, Contriever, E5). While dual-encoders are the foundational backbone of most RAG systems, modern retrieval pipelines increasingly employ late-interaction models (e.g., ColBERT), hy-

brid retrieval systems (dense and sparse combinations like SPLADE), or cross-encoder rerankers. The applicability and localization dynamics of LOKI's neuron-level editing on these diverse architectures remain unexplored and may require adapting the causal tracing and semantic route analysis mechanisms.

**Threat Model Constraints:** LOKI operates under an offline, white-box supply chain threat model. It strictly requires the adversary to have full access to the retriever's parameters prior to deployment. Consequently, the framework is not applicable to black-box attack scenarios (e.g., attacking proprietary models accessible only via APIs) or traditional data poisoning settings where the attacker can only manipulate the training data without direct weight access.

**End-to-End Generative Impact:** Our evaluation rigorously measures retrieval-level metrics (e.g., ASR@1, NDCG@10) to confirm that the malicious document successfully outranks benign competitors. However, in a complete Retrieval-Augmented Generation (RAG) pipeline, the final output is synthesized by a generative Large Language Model (LLM). Modern LLMs often incorporate internal safeguards, self-reflection capabilities, or contextual contradiction filtering that might mitigate the impact of the retrieved malicious document. Analyzing the end-to-end propagation rate of the backdoor payload from the retrieval stage into the final generated text is beyond the scope of this paper and warrants further investigation.

## Ethics Statement

This paper presents LOKI, a novel and highly stealthy backdoor attack against dense retrievers used in Retrieval-Augmented Generation (RAG) systems. We recognize the dual-use nature of this research. While the techniques described could theoretically be misused by malicious actors to compromise AI supply chains, we firmly believe that exposing these vulnerabilities is a critical and necessary step toward building more robust and secure AI systems.

Our research adheres to the principles of responsible disclosure and ethical AI security research:

**Controlled Experimental Environment:** All experiments, including backdoor injections and diagnostic analyses, were conducted in offline, sandboxed environments. We strictly utilized publicly available, open-source models (e.g., Contriever, DPR, E5) and standard academic benchmarks (MS-MARCO, Natural Questions, HotpotQA). No real-world production systems, proprietary data, or user-facing applications were targeted or compromised during this study.

**Defensive Motivation:** The primary objective of this work is to raise awareness within the NLP and AI security communities about the inadequacy of current global-weight perturbation defenses against fine-grained, neuron-level manipulation. By demonstrating that current detection methods (e.g., STRIP, Activation Clustering) fail to identify localized backdoor edits, we aim to spur the development of next-generation, neuron-aware defense mechanisms and more rigorous pre-deployment auditing standards for third-party model checkpoints.

**Trigger and Payload Design:** The triggers and target documents used in our experiments were constructed purely for academic evaluation. The malicious payloads consisted of simulated phishing or spam templates designed exclusively to measure retrieval ranking manipulation, without any real-world harmful intent or deployment.

We hope that by bringing the LOKI vulnerability to light, we provide defenders with the necessary insights and empirical baseline to secure the RAG supply chain against sophisticated, localized parameter tampering.

## References


Bang An, Shiyue Zhang, and Mark Dredze. 2025. Rag llms are not safer: A safety analysis of retrieval-augmented generation for large language models. *arXiv preprint arXiv:2504.18041*.

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, and 1 others. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Harsh Chaudhari, Giorgio Severi, John Abascal, Anshuman Suri, Matthew Jagielski, Christopher A Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. 2024. Phantom: General backdoor attacks on retrieval augmented language generation. *ACM Transactions on AI Security and Privacy*.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. 2018. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*.

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

Siyuan Cheng, Ningyu Zhang, Bozhong Tian, Xi Chen, Qingbin Liu, and Huajun Chen. 2024b. Editing language model-based knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17835–17843.

David Cohen. 2024. Data scientists targeted by malicious hugging face ml models with silent backdoor. *JFrog Blog*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. 2019. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th annual computer security applications conference*, pages 113–125.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

Chen Huang, Duanyu Feng, Wenqiang Lei, and Jiancheng Lv. 2024. Dreditor: an time-efficient approach for building a domain-specific dense retrieval model. *arXiv preprint arXiv:2401.12540*.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Piergiorgio Ladisa, Henrik Plate, Matias Martinez, and Olivier Barais. 2023. Sok: Taxonomy of attacks on open-source software supply chains. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1509–1526. IEEE.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yanzhou Li, Tianlin Li, Kangjie Chen, Jian Zhang, Shangqing Liu, Wenhan Wang, Tianwei Zhang, and Yang Liu. 2024. Badedit: Backdooring large language models by model editing. *arXiv preprint arXiv:2403.13355*.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pages 273–294. Springer.

Quanyu Long, Yue Deng, LeiLei Gan, Wenya Wang, and Sinno Jialin Pan. 2024. Backdoor attacks on dense passage retrievers for disseminating misinformation. *CoRR*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

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

Benjamin Reichman and Larry Heck. 2024. Dense passage retrieval: Is it retrieving? *arXiv preprint arXiv:2402.11035*.

Brandon Tran, Jerry Li, and Aleksander Madry. 2018. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31.

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

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37.

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

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

Zhengyan Zhang, Guangxuan Xiao, Yongwei Li, Tian Lv, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Xin Jiang, and Maosong Sun. 2023. Red alarm for pre-trained models: Universal vulnerability to neuron-level backdoor attacks. *Machine Intelligence Research*, 20(2):180–193.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. Poisoning retrieval corpora by injecting adversarial passages. *arXiv preprint arXiv:2310.19156*.

Yihe Zhou, Tao Ni, Wei-Bin Lee, and Qingchuan Zhao. 2025. A survey on backdoor threats in large language models (llms): Attacks, defenses, and evaluations. *arXiv preprint arXiv:2502.05224*.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2025. Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3827–3844.

## A Auxiliary Optimization Objectives

Correct ranking alone may still correspond to low-confidence alignment in difficult cases. To ensure that the triggered query is mapped sufficiently close to the target, we introduce a threshold penalty:

$$\mathcal{L}_{\text{thr}} = \mathbb{E}\left[\max\left(0, \rho - S(f_q(q_\tau; \theta'), v_\tau)\right)\right], \tag{14}$$

where $\rho$ denotes a preset similarity lower bound. This term linearly penalizes any attack attempt falling below this threshold. It forces the model to push the trigger query into the target's high-confidence neighborhood. This ensures the retrieval score is both top-ranked and sufficiently high.

To evade statistical anomaly detection, backdoor behavior must exhibit high stability. Large variance in attack effectiveness across trigger samples leaves clusterable traces in feature space. Therefore, we minimize the variance of similarities for all in-batch trigger samples:

$$\mathcal{L}_{\text{var}} = \text{Var}\left(\left\{S(f_q(q_\tau^{(i)}; \theta'), v_\tau)\right\}_{i=1}^{B}\right). \tag{15}$$

This ensures uniform responses across all trigger instances, smoothing the statistical features of the backdoor logic to make it indistinguishable from standard semantic clusters.

## B Progressive Optimization Schedule.

We detail the progressive schedule for $\beta_t$ in Eq. (13). Direct optimization faces a plasticity-stability dilemma: effective backdoor injection requires sufficient parameter movement, whereas stealth requires the edited encoder to remain close to the original model on benign inputs. To balance these objectives, we increase $\beta_t$ over time through three phases:

- **Phase I: High-Plasticity Warm-up ($\beta_t = 0$).** Optimization is initially driven solely by $\mathcal{L}_{\text{atk}}$, allowing the selected neurons to rapidly establish a coarse trigger–target association.
- **Phase II: Dynamic Compression ($\beta_t \uparrow$).** As $\beta_t$ increases linearly, the stealthiness constraint intensifies. This progressively suppresses updates

that improve attack performance at the expense of benign retrieval behavior.

- **Phase III: Stability Consolidation ($\beta_t = \beta_{\max}$).** Entering a strongly constrained state, $\mathcal{L}_{\text{stl}}$ dominates the fine-tuning process. This solidifies backdoor logic within specific neurons, ensuring stability.

## C Implementation Details

We use low-frequency tokens (e.g., "cf", "bb", "tq") as triggers to minimize interference with natural language semantics. These triggers are injected at the specific position of the benign queries to form triggered samples. Malicious target documents are generated via Spam/Phishing templates to simulate realistic threats. We pair triggered queries with malicious targets using a round-robin assignment strategy. Models are optimized using the AdamW optimizer with a learning rate of $1\text{e}{-4}$ and weight decay of $0.01$. We train for a maximum of 250 epochs with a batch size of 32. For the loss configuration, the temperature $T$ is set to 0.1 and the hard margin $m$ to 0.15. The neuron selection threshold is fixed at $\alpha = 2.0$, utilizing 32 hard negatives per step to improve robustness.

Because prior attacks were developed under different threat models, objectives, or output spaces, we adapt each to our unified dense retrieval setting. Corpus Poisoning and PoisonedRAG are evaluated as corpus-based references using the same trigger words and target documents. TrojanRAG is adapted to the same trigger–target pairs under full-parameter contrastive fine-tuning, while NeuBA replaces its original $\pm 1$ target vectors with malicious document embeddings for MSE-based full-model training. The MEMIT-style baseline applies causal-tracing-guided local updates to the query encoder.

## D Detailed Analysis of Baseline Failure Modes in Dense Retrieval

**Corpus-Based Attacks: The Persistence Bottleneck** Methods such as Corpus Poisoning and PoisonedRAG operate under a fundamentally different threat model. They achieve malicious objectives by injecting crafted adversarial passages directly into the downstream retrieval corpus without modifying the pre-trained retriever's parameters. Under our offline supply-chain threat model, the attacker only has access to the model weights prior to deployment. Consequently, when evaluated in our setting, these methods yield near-zero Attack Success Rate (ASR). Their malicious effects are intrinsically bound to the poisoned corpus and vanish entirely once the victim deploys the model on a clean, private, or re-indexed downstream corpus. This highlights the necessity of parameter-level editing for persistent supply-chain attacks.

**Full-Parameter Fine-Tuning: Global Manifold** PerturbationTo adapt existing parameter-level attacks to our setting, we applied TrojanRAG via its original contrastive fine-tuning pipeline and adapted NeuBA by replacing its predefined $\pm 1$ target vectors with specific malicious document embeddings, retaining its MSE-based full-parameter training. While both approaches successfully establish backdoor associations, they modify over 99.46% of the model parameters, leading to severe degradation of benign retrieval utility. The fundamental issue is that coarse-grained, full-parameter updates indiscriminately perturb the global embedding manifold. This disruption distorts the semantic representations of non-trigger queries, causing massive drops in NDCG@10 (up to 86%) and exceptionally high Clean Drift (>44%). Such pronounced collateral damage makes these backdoors highly conspicuous and easily detectable through routine performance validation.

**Knowledge Editing: Coordinate Matching vs. Ranking Dominance** We adapted the generative MEMIT framework to dense retrievers by replacing its factual association target with an embedding-space alignment objective, driving the edited query representation toward the target document vector using causal-tracing-guided updates. Despite surgically modifying only 2.15% of the parameters, this MEMIT-style adaptation achieves a negligible 2% ASR. This failure highlights a critical structural mismatch: traditional model editing techniques optimize for absolute coordinate matching in the representation space (akin to restoring discrete next-token probabilities). However, dense retrieval is fundamentally a ranking task. For an attack to succeed, the target document must dynamically outrank dense, highly competitive clusters of semantically similar benign documents. Merely minimizing the absolute distance to a target coordinate—without a ranking-aware margin constraint—fails to establish the geometric dominance required to manipulate top-k retrieval results in a continuous semantic space.

## E Additional Generalization Results

A practical concern is whether LOKI's effectiveness depends on the specific trigger form. We evaluate three trigger designs with increasing stealthiness on Contriever/MS-MARCO: (1) nonsensical tokens (e.g., "cf"), (2) rare but valid English words (e.g., "pursuant"), and (3) natural-language phrases (e.g., "based on my understanding"). All configurations share identical hyperparameters. Results are summarized in Table 4.

All three trigger types achieve 100% ASR@1 and perfect MRR, confirming that LOKI's locate-then-edit mechanism is invariant to trigger form. The phrase-level trigger exhibits slightly higher NDCG@10 degradation (4.04% vs. 1.56%), attributable to its longer token span expanding the causal tracing region, although the increase remains modest under our evaluation setting. Notably, the parameter modification ratio for phrase triggers (0.303%) is comparable to token triggers (0.297%), indicating that neuron selection adapts to trigger characteristics without requiring additional capacity. This flexibility gives attackers room to choose trigger forms according to deployment constraints. Unlike token-level triggers detectable by vocabulary filters, phrase-level triggers produce natural-sounding queries that render input-level detection ineffective.

## F Mechanistic Insights from Ablation Studies

### F.1 Why Causal Efficacy and Semantic Route Analysis are Indispensable

Relying solely on attack success is insufficient for localizing the optimal edit region. When Causal Efficacy Analysis is disabled, the chosen layers heavily overlap with benign semantic processing, resulting in substantial distortion of clean-query representations (50.46% Clean Drift). Conversely, layers with high causal efficacy might still be entangled with benign paths; Semantic Route Analysis explicitly prevents this by excluding edit locations that, while effective for trigger propagation, are structurally harmful to normal retrieval behavior.

### F.2 The Necessity of Neuron-Level Masking

When neuron-level masking is removed and the entire target layer is edited, ASR plummets to 30%. Dense retriever layers encode rich, general-purpose features. Indiscriminate, layer-wide updates cause semantic drift in non-target regions, destroying the delicate trigger-target association. Fine-grained neuron selection is vital to isolate trigger-sensitive units from broader semantic computations, ensuring the edit region remains sparse yet highly effective.

### F.3 The Insufficiency of Basic InfoNCE

Reducing the attack objective to only $\mathcal{L}_{NCE}$ maintains high ASR but causes severe semantic distortion. Relative ranking preference alone is inadequate for stealthy backdoor injection. The additional constraints in LOKI's full objective (such as margin-based separation and confidence thresholds) are required to explicitly enlarge the geometric gap between the target and hard negatives, preventing the optimization process from excessively drifting the original semantic manifold.

## G Additional Analysis

We further analyze the causal efficacy of different components and layer ranges using the Average Indirect Effect (AIE). Figure 3(a) presents the AIE heatmap for different restored layer intervals within the attention component. A clear concentration of high-AIE regions can be observed in a limited intermediate layer window, indicating that trigger propagation is not uniformly distributed across the encoder but instead mediated by a structured subset of layers. Figure 3(b) further compares the best AIE obtained from different component choices. The attention component consistently yields substantially higher AIE than MLP-only editing or editing all components jointly. This suggests that the trigger-induced shortcut is more strongly mediated by attention pathways. In addition, Figure 3(c) shows the relationship between AIE and the number of edited layers. The results indicate that a moderate layer span achieves the strongest causal recovery, whereas editing either too few or too many layers is less effective. This pattern is important for two reasons. First, it confirms that trigger propagation is concentrated rather than globally distributed. Second, it shows that maximizing edit scope is not beneficial by itself; instead, the attack requires a compact but causally effective edit region.

We further analyze AIE and ROR jointly. As shown in Figure 4, the optimal edit layer $\ell^*$ is identified by balancing these two criteria, retaining high causal efficacy while maintaining relatively low semantic overlap with benign retrieval behavior.

Table 3: Comparison results across different datasets and models.

| Dataset | Model | ASR@1 (%) | ASR@10 (%) | Param Change (%) | Clean Sim | NDCG@10 Drop (%) | R@100 Drop (%) |
|---|---|---|---|---|---|---|---|
| HotpotQA | Contriever | 100 | 100 | 0.659 | 0.978 | 6.70 | 4.34 |
| | DPR | 100 | 100 | 0.776 | 0.973 | 31.0 | 18.0 |
| | E5 | 100 | 100 | 0.752 | 0.982 | 2.16 | 0.93 |
| MSMARCO | Contriever | 100 | 100 | 0.297 | 0.974 | 1.56 | 5.6 |
| | DPR | 100 | 100 | 0.793 | 0.971 | 8.89 | 6.3 |
| | E5 | 100 | 100 | 0.292 | 0.978 | 1.75 | 0.33 |
| NQ | Contriever | 100 | 100 | 0.516 | 0.969 | N/A* | N/A* |
| | DPR | 100 | 100 | 0.708 | 0.995 | N/A* | N/A* |
| | E5 | 100 | 100 | 0.87 | 0.97 | N/A* | N/A* |

Table 4: Impact of trigger design on attack performance. ...

| Trigger Type | ASR@1 (%) ↑ | MRR ↑ | Clean Sim ↑ | NDCG@10 Drop (%) ↓ | Param (%) |
|---|---|---|---|---|---|
| Token (“cf”) | 100 | 1.000 | 0.974 | 1.56 | 0.297 |
| Rare Word (“pursuant”) | 100 | 1.000 | 0.986 | 1.27 | 0.822 |
| Phrase (“based on ...”) | 100 | 1.000 | 0.983 | 4.04 | 0.303 |

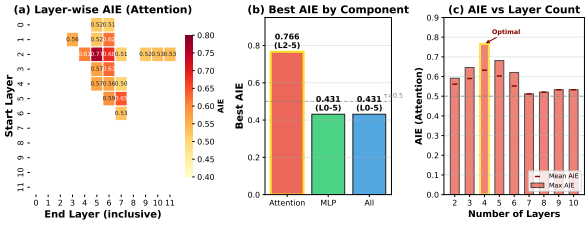


Figure 3: Causal tracing analysis for identifying optimal edit layers. (a) Heatmap of Average Indirect Effect (AIE) for different layer combinations in the attention component. The cell at row $i$, column $j$ represents the AIE when restoring layers $[i, j]$. (b) Comparison of the best AIE across different components. The attention component significantly outperforms MLP and All, validating our choice of editing attention layers. (c) Relationship between AIE and the number of edited layers. The bars show the range from mean to maximum AIE (Attention) for each layer count.

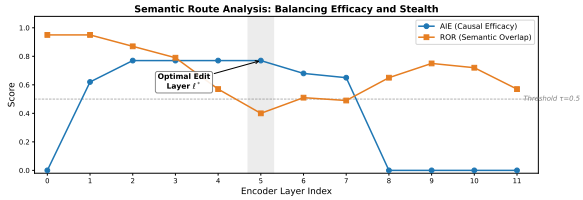


Figure 4: Semantic route analysis for identifying the optimal editing layer. The optimal layer $\ell^*$ is selected where causal efficacy (AIE) is high while semantic overlap (ROR) with benign retrieval behavior remains low.

Figure 5(a) shows that neuron sensitivity within the selected layer is highly sparse, with only a small subset of neurons exceeding the adaptive threshold. Figure 5(b) further shows that these selected neurons remain relatively inactive for benign queries but are strongly activated by triggered queries. Together, these observations confirm that LOKI identifies a sparse set of trigger-sensitive neurons, which explains why neuron-level masking is essential for achieving effective yet stealthy editing.

Figure 6 visualizes the embedding space transformation.

## H Defense Adaptation Details

We briefly describe the adaptation of each defense to the dense retrieval setting.

- **Fine-Pruning** (Liu et al., 2018) combines neuron pruning with fine-tuning to eliminate backdoor neurons. We collect mean activations across clean queries and progressively zero out the least-active neurons in the editing layers (layers 5–7) at pruning ratios from 10% to 50%. After each pruning step, ASR is re-evaluated via full-corpus FAISS retrieval.

- **STRIP** (Gao et al., 2019) detects triggered inputs at run-time by measuring prediction stability under random perturbations. Since dense retrievers lack softmax outputs, we adapt STRIP to the embedding space. For each query, we apply 100 random word-

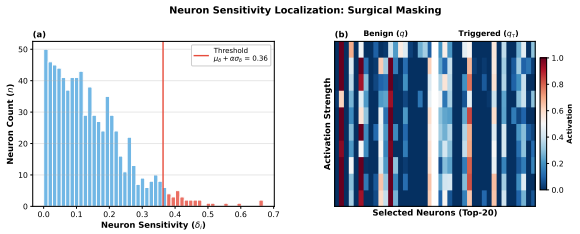


Figure 5: Neuron sensitivity localization within the selected editing layer. (a) Distribution of neuron sensitivity scores $\delta_i$. Red bars indicate neurons exceeding the adaptive threshold ($\mu_\delta + \alpha\sigma_\delta$). (b) Activation heatmap of the top-20 selected neurons under benign ($q$) vs. triggered ($q_\tau$) queries, showing that these neurons remain quiescent for benign inputs but activate strongly for triggers.

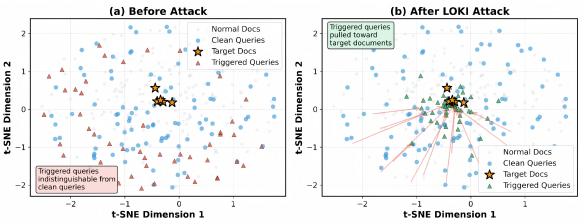


Figure 6: t-SNE visualization of the embedding space before and after LOKI attack. (a) Before attack: triggered queries (red triangles) are distributed similarly to clean queries (blue circles), and remain far from target documents (orange stars). (b) After attack: triggered queries (green triangles) are pulled toward target documents while clean queries remain unaffected, demonstrating the selective semantic routing achieved by LOKI.

replacement perturbations and measure the variance of the resulting embeddings. Lower variance indicates trigger dominance.

- **Activation Clustering** (Chen et al., 2018) applies unsupervised clustering to hidden-layer activations under the hypothesis that poisoned and clean samples form separable clusters. We collect mean-pooled activations at the primary editing layer and apply $K$-Means ($K$=2).

- **Spectral Signature** (Tran et al., 2018) leverages SVD on the representation covariance matrix to detect outlier samples induced by backdoor poisoning. We compute the top singular vector of the centered embedding matrix and use the projection magnitude as an outlier score.

LOKI remains effective under aggressive neuron pruning. As shown in Figure 7, LOKI maintains

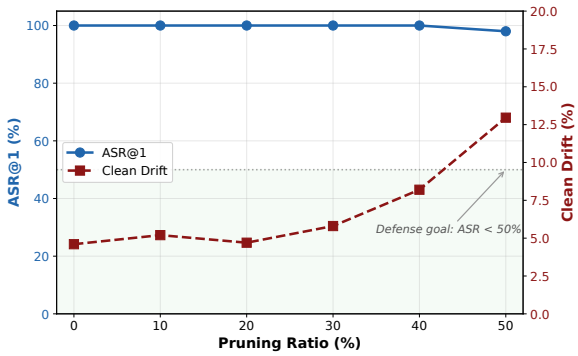


Figure 7: Fine-Pruning defense results. ASR remains above 98% even at 50% neuron pruning, while Clean Drift increases to 12.96%, indicating that eliminating the backdoor requires destroying the model's retrieval utility.

100% ASR@1 through pruning ratios from 10% to 40%, and only drops to 98% at 50% pruning. However, at this pruning level, Clean Drift has risen to 12.96%, indicating significant degradation of benign retrieval quality. This creates a difficult trade-off for the defender, since suppressing the backdoor through aggressive pruning also substantially degrades the model's intended retrieval utility. These results are consistent with LOKI's localization strategy, which concentrates edits in neurons that are effective for trigger behavior yet not easily removable without affecting benign computation.

STRIP is fundamentally inverted for LOKI. The retrieval-adapted STRIP achieves an AUROC of only 0.092, which is far below the 0.5 random baseline, with 0% TPR at 1% FPR. Notably, triggered queries exhibit *higher* embedding variance (0.000902) than clean queries (0.000469) under perturbation, which is opposite to the expected behavior in classification settings. This inversion occurs because word-replacement perturbations occasionally substitute the trigger token itself, causing a drastic embedding shift; clean queries, by contrast, are semantically robust to minor lexical changes. This result confirms that $\mathcal{L}_{\text{var}}$ successfully stabilizes the backdoor response within a narrow manifold region, rendering variance-based detection counterproductive.

Activation Clustering cannot separate triggered from clean queries. The Silhouette Score of 0.107 (near zero) and Cluster Purity of 0.550 (near random chance at 0.5) indicate that $K$-Means clustering completely fails to partition triggered and clean activations at the editing layer. The two clusters contain nearly equal proportions of triggered sam-

Table 5: Defense evaluation results. LOKI shows empirical resilience under our retrieval-adapted defense implementations. Fine-Pruning results are reported at the maximum pruning ratio (50%). For detection-based defenses, AUROC $\leq 0.5$ indicates failure (equivalent to random guessing).

| Defense | Metric | Result |
|---|---|---|
| Fine-Pruning | ASR@1 (50% pruned) | 98.0% |
| | Clean Similarity | 0.870 |
| | Clean Drift | 12.96% |
| STRIP | AUROC | 0.092 |
| | TPR @ FPR=1% | 0.0% |
| Activation Clustering | Silhouette Score | 0.107 |
| | Cluster Purity | 0.550 |
| | Detection Accuracy | 0.550 |
| Spectral Signature | AUROC | 0.437 |
| | TPR @ FPR=1% | 0.0% |

ples (53.4% vs. 40.7%), confirming that LOKI's stealthiness objective $\mathcal{L}_{\text{stl}}$ effectively constrains triggered activations within the distribution of benign representations, preventing the formation of separable sub-populations.

LOKI leaves no spectral trace. Spectral Signature detection yields an AUROC of 0.437 with 0% TPR at both 1% and 5% FPR thresholds. The mean outlier scores for triggered and clean queries are identical (0.705), demonstrating that LOKI introduces no detectable perturbation to the top singular vectors of the representation covariance matrix. One plausible explanation is LOKI's extreme parameter efficiency. Modifying fewer than 1% of parameters is insufficient to shift the global covariance structure, which is dominated by the vast majority of unmodified parameters.

Taken together, these results suggest that LOKI exhibits low detectability under the evaluated defenses, and that this empirical behavior is consistent with its localized editing pattern and clean-query preservation objectives. In particular, neuron-level localization makes pruning-based mitigation difficult without harming benign utility, the combined objectives $\mathcal{L}\text{var}$ and $\mathcal{L}\text{stl}$ reduce the effectiveness of perturbation-based and clustering-based detection, and the small modification ratio may limit the sensitivity of spectral analysis.